IJCSIS Vol. 14 Special issue 1, February 2016
ISSN 1947-5500

International Journal of Computer Science & Information Security

© IJCSIS PUBLICATION 2016
Pennsylvania, USA

Cornell University Library

Cogprints

Google scholar

.docstoc
find and share professional documents

ScientificCommons

View my documents on Scribd

BASE
Bielefeld Academic Search Engine

SCIRUS
search engine for science

SciRate.com

CiteSeerx beta

dblp.uni-trier.de
Computer Science Bibliography

Q·Sensei BETA

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

EBSCO HOST

ProQuest

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2016 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Editorial Board

It is our great pleasure to present to you the first ***Special Issue 2016*** *of the* ***International Journal of Computer Science and Information Security*** *(IJCSIS). The journal offers survey and review articles from experts in the field, promoting insight and understanding of the state of the art, and trends in Special issue on "Computing Applications and Data Mining". The contents include original research and innovative applications from all parts of the world. While the journal presents mostly previously unpublished materials, selected conference papers with exceptional merit are also published, at the discretion of the editors. The main objective is to disseminate new knowledge and latest research for the benefit of all, ranging from academia and professional communities to industry professionals.*

According to Google Scholar, up to now papers published in ***IJCSIS*** *have been cited over 5668 times and the number is quickly increasing. This statistics shows that* ***IJCSIS*** *has established the first step to be an international and prestigious journal in the field of Computer Science and Information Security. IJCSIS archives all publications in major academic/scientific databases and is now indexed by the following International agencies and institutions: Google Scholar, CiteSeerX, Cornell's University Library, Ei Compendex, ISI Scopus, DBLP, DOAJ, ProQuest, Thomson Reuters, ArXiv, ResearchGate, Academia.edu and EBSCO among others.*

The editorial board is pleased to present the ***Special Issue February 2016 issue****. We thank and congratulate the IJCSIS team, associate editors, and reviewers for their dedicated services to review and recommend high quality papers for publication. In particular, we would like to thank distinguished international authors for submitting their papers to IJCSIS and researchers for continued support by citing papers published in IJCSIS. Without their sustained and unselfish commitments, IJCSIS would not have achieved its current premier status.*

"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication."

For further questions or other suggestions please do not hesitate to contact us at ***ijcsiseditor@gmail.com****.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 14, Special Issue 1 (S1), February 2016 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

Bibliographic Information
ISSN: 1947-5500
Monthly publication (Regular Special Issues)
Commenced Publication since May 2009

Editorial / Paper Submissions:
IJCSIS Managing Editor
(ijcsiseditor@gmail.com)
Pennsylvania, USA
Tel: +1 412 390 5159

IJCSIS EDITORIAL BOARD

ISSN 1947 5500 Copyright © IJCSIS, USA.

TABLE OF CONTENTS

1. Paper 01021601: A New Efficient Approach for Renewable Energy Ontology (pp. 1-8)

Awwad Ahmad (1), Kayed Ahmad (2)*
(1) Faculty of Information Technology, University of Petra, Amman, Jordan.
*(*2) Faculty of Information Tech., Middle East University, MEU, Amman, Jordan.*

Abstract: All over the world renewable energy implementations and applications are becoming a very crucial issue to their successful. Taking in consideration that a specific piece of information, service feedback, or product from an electronic provider is trustable and reliable may be a difficult task sometimes. As we know that World Wide Web (WWW) is an open environment in which it allows any person to distribute huge amounts of information. The accuracy or reliability of such information, to some degree, is unknown, and therefore cannot be trusted. In this research paper, we claim and argue that using ontology may form a useful tool to find the best renewable energy provider. The contribution of this paper is to develop ontology concepts for measuring such "goodness". Common and frequent concepts from five popular and trusted online renewable energy providers were extracted, distinguished, and then checked against nine other online providers. These providers are also judged by experts who are renewable energy specialists. The results discussed and argued in this paper have shown that the proposed approach has achieved high matching score to the experts' judgments.

Keywords: Ontology; Semantic, Reliability; Accuracy, Renewable Energy.

2. Paper 01021602: Using Graphs to Visualize Overlapping Sets (pp. 9-18)

Abdalmunam Abdalla (1), Mehmet Koyuturk (2), Abdelsalam M. Maatuk (3), Alfaroq O. Mohammed (4)
(1,4) Department of Computer Science, Omer AL-Mukhtar University, Libya
(3) Faculty of Information Technology, Benghazi University, Libya
(2) Faculty of Engineering, Case Western Reserve University, USA

Abstract: The large volume of data available in many domains and the need to analyze the data to extract useful information from it has lead to the need of visualization techniques to get information about the data at a glance. Visual inspection is useful in providing fast and abstract information about datasets to guide the researchers in choosing the suitable approach to process the data. Recently, there have been notable advances in graph visualization; however, visualizing sets still needs more attention. In this paper a method is proposed to visualize overlapping sets so that the underlying hierarchy and relations of the sets can be easily understood by visual inspection. This approach utilizes the graph representation of the sets to aid the drawing process. Using the spectral decomposition of the graph derived from the sets, we developed algorithms to compute the best coordinates for the items of the sets and plot them on the Euclidean plane. The method has been tested on both real and synthetic datasets to investigate its performance.

Keywords: Sets, sets visualization, overlapping sets, visualization, sets drawing.

3. Paper 01021603: Shortest Path of Metrics TYPE (pp. 19-22)

Ahmad Tayyar, Jerash University, Jordan

Abstract — A geodesic is the real world analog of a straight line. Where a straight line on a flat piece of paper minimizes the distance between two points, a geodesic minimizes the distance between two points on any surface; be it flat or not. Supposing that we have a surface in space given by the equation z = f (x, y).The search for a geodesic line on this surface, or more generally in the plan provided by an arbitrary metric, may be made by solving the coupled differential second order equations of Euler-Lagrange system. More precisely, the search of the shortest path connecting two given points may be made by solving that system for a specific initial velocity.

In this paper we determine the geodesic lines corresponding the metric of type g = (dx2 + dy2) for f (x, y) defines positive.

Starting from a metric of this type, we determine the Euler-Lagrange system correspondence; its solutions are geodesics. We designed geodesics and the shortest path for the given metric and a specific function f (x, y).

We will need to determine the appropriate initial velocity for the system's numerical resolution of two differential equations of second order. Therefore, we are providing a suitable method for this.

Keywords— Euler-Lagrange, geodesics, metric, shortest path.

4. Paper 01021604: An Innovative Imputation and Classification Approach for Accurate Disease Prediction (pp. 23-31)

Yelipe UshaRani, Department of Information Technology, VNR VJIET, Hyderabad, INDIA
Dr. P. Sammulal, Dept.of Computer Science and Engineering, JNT University, Karimnagar, INDIA

Abstract — Imputation of missing attribute values in medical datasets for extracting hidden knowledge from medical datasets is an interesting research topic of interest which is very challenging. One cannot eliminate missing values in medical records. The reason may be because some tests may not been conducted as they are cost effective, values missed when conducting clinical trials, values may not have been recorded to name some of the reasons. Data mining researchers have been proposing various approaches to find and impute missing values to increase classification accuracies so that disease may be predicted accurately. In this paper, we propose a novel imputation approach for imputation of missing values and performing classification after fixing missing values. The approach is based on clustering concept and aims at dimensionality reduction of the records. The case study discussed shows that missing values can be fixed and imputed efficiently by achieving dimensionality reduction. The importance of proposed approach for classification is visible in the case study which assigns single class label in contrary to multi-label assignment if dimensionality reduction is not performed.

Keywords— imputation; missing values; prediction; nearest neighbor, cluster, medical records, dimensionality reduction

5. Paper 01021605: A Study of Adopting Cloud Computing from Enterprise Perspective using Delone and Mclean IS Success Model (pp. 32-38)

Bassam Al-Shargabi, Omar Sabri, Isra University, Amman-Jordan

Abstract: Nowadays, Cloud Computing is the new promising technology that enable sharing resources between different enterprises through Internet in an on-demand manner. Many enterprises are moving toward adopting cloud computing services to gain the benefits of cost reduction of such services. Thus , many enterprises are facing greater obstacles for adapting this new technology, In this paper, the DeLone and McLean successes model is used to assess and evaluate some components that need to be considered by an enterprise when making the decision of adopting cloud computing. The enterprise will be able to identify its weakness and strength for each factor, and then build and prepare plan that can help them to make appropriate decision toward a successful adoption of Cloud Computing.

Keywords: Cloud Computing; Information Technology; Software as a Service; Infrastructure as a Service; Platform as a Service

6. Paper 01021606: Clustering and Classification of Text Documents Using Improved Similarity Measure (pp. 39-54)

G. SureshReddy (1), T.V. Rajinikanth (2), A. AnandaRao (3)
(1) Department of Information Technology, VNR VJIET, Hyderabad, India
(2) Department of Computer Science and Engineering, SNIST, Hyderabad, India
(3) Department of Computer Science and Engineering, JNTU University, Anantapur, India

Abstract: Dimensionality reduction is very challenging and important in text mining. We need to know which features be retained what to be and It helps in reducing the processing overhead when performing text classification and text clustering. Another concern in text clustering and text classification is the similarity measure which we choose to find the similarity degree between any two text documents. In this paper, we work towards text clustering and text classification by addressing dimensionality reduction using SVD followed by the use of the proposed similarity measure which is an improved version of our previous measure (25, 31). This proposed measure is used for supervised and un-supervised learning. The proposed distance measure overcomes the disadvantages of the existing measures (10).

Keywords: Feature Vector, Similarity, Feature Set, Commonality

7. Paper 01021607: QoS Web Service Security Dynamic Intruder Detection System for HTTP SSL services (pp. 55-60)

M. Swami Das (1), A. Govardhan (2), D. Vijaya lakshmi (3)
(1) Assoc. Professor, CSE, MREC
(2) Professor, SIT, JNTU Hyderabad
(3) Professor, Dept. of CSE, MGIT Hyderabad, India

Abstract: Web services are expected to play significant role for message communications over internet applications. Most of the future work is web security. Online shopping and web services are increasing at rapid rate. In this paper we presented the fundamental concepts related to Network security, web security threats. QoS web service security intrusion detection is important concern in network communications and firewalls security; we discussed various issues and challenges related to web security. The fundamental concepts network security XML firewall, XML networks. We proposed a novel Dynamic Intruder Detection System (DIDA) is safe guard against SSL secured transactions over message communications in intermediate routers that enable services to sender and receiver use Secured Session Layer protocol messages. This can be into three stages 1) Sensor 2) Analyzer and 3)User Interface.

Keywords: Web Security, QoS web service, HTTP, Intruder detection, Secure Socket Layer, Network Security

8. Paper 01021608: Does Software Structures Quality Improve over Software Evolution? Evidences from Open-Source Projects (pp. 61-75)

Mamdouh Alenezi and Mohammad Zarour College of Computer & Information Sciences Prince Sultan University, Riyadh 11586 Saudi Arabia

Abstract - Throughout the software evolution, several maintenance actions such as adding new features, fixing problems, improving the design might negatively or positively affect the software design quality. Quality degradation, if not handled in the right time, can accumulate and cause serious problems for future maintenance effort. Several researchers considered modularity as one of the success factors of Open Source Software (OSS) Projects. The modularity of these systems is influenced by some software metrics such as size, complexity, cohesion, and coupling. In this work, we study the modularity evolution of four open-source systems by answering two main research questions namely: what measures can be used to measure the modularity level of software and secondly, did the modularity level for the selected open source software improves over time. By investigating the modularity measures, we have identified the main measures that can be used to measure software modularity. Based on our analysis, the modularity of these two systems is not improving over time. However, the defect density is improving over time.

9. Paper 01021609: Intrusion Detection – A Text Mining Based Approach (pp. 76-88)

Gunupudi RajeshKumar (1), N. Mangathayaru (2), G. Narsimha (3)
(1,2) Faculty of Information Technology, VNR VJIET, India
(3) Faculty of Computer Science and Engineering, JNT University, Jagityal, India

Abstract: Intrusion Detection is one of major threats for organization. The approach of intrusion detection using text processing has been one of research interests which is gaining significant importance from researchers. In text mining based approach for intrusion detection, system calls serve as source for mining and predicting possibility of intrusion or attack. When an application runs, there might be several system calls which are initiated in the background. These system calls form the strong basis and the deciding factor for intrusion detection. In this paper, we mainly discuss the approach for intrusion detection by designing a distance measure which is designed by taking into consideration the conventional Gaussian function and modified to suit the need for similarity function. A Framework for intrusion detection is also discussed as part of this research.

Keywords: system calls, intrusion, prediction, classification, kernel measures

A New Efficient Approach for Renewable Energy Ontology

Awwad Ahmad[1], Kayed Ahmad[2 *]

[1]*Faculty of Information Technology, University of Petra, Amman, Jordan.*

[2*]*Faculty of Information Tech., Middle East University, MEU, Amman, Jordan.*

Abstract: All over the world renewable energy implementations and applications are becoming a very crucial issue to their successful. Taking in consideration that a specific piece of information, service feedback, or product from an electronic provider is trustable and reliable may be a difficult task sometimes. As we know that World Wide Web (WWW) is an open environment in which it allows any person to distribute huge amounts of information. The accuracy or reliability of such information, to some degree, is unknown, and therefore cannot be trusted. In this research paper, we claim and argue that using ontology may form a useful tool to find the best renewable energy provider. The contribution of this paper is to develop ontology concepts for measuring such "goodness". Common and frequent concepts from five popular and trusted online renewable energy providers were extracted, distinguished, and then checked against nine other online providers. These providers are also judged by experts who are renewable energy specialists. The results discussed and argued in this paper have shown that the proposed approach has achieved high matching score to the experts' judgments.

Keywords*:* Ontology; Semantic, Reliability; Accuracy, Renewable Energy.

1. Introduction

These days more researchers and people using the (WWW) to look for and get some specific information, to arrange for their holidays or scientific trips, to even to buy or sell ANTICQUES, to do financial transactions, or even to play specific internet games. However, these people and researchers facing problems with the credibility and trustability of the sites that they are using [1]. A big number of the loaded web sites still being uploaded to the internet. The quality of information that different users find on the net from new or unpopular sites is still questionable and unambiguous. As a result, web designers and developers face more pressure to enhance the credibility and accuracy of their web sites [2, 3]. The problem still that there are no well-known standards criteria and measures for web site credibility, so designing and implementing a website is becoming even now more and more difficult than any time before.

Simple, professional, and friendly designs, as well as customers' testimonials and photo galleries used to be ways to prove credibility and reliability of a web site. Unfortunately, these are currently being used by both trustworthy and untrustworthy sites. More studies are made to discuss the different factors and elements of web sites that affect people’s perception to credibility. These elements have been divided into seven types. The five types that affect positively credibility awareness are “real-world feel”, “ease of use”, “expertise”, trustworthiness”, reliability and “tailoring”, on the other hand two other types that affect in negative way the credibility are “commercial implications” and “amateurism” [4]. We aims in this research to discuss and study the credibility of web sites that concerns the Renewable Energy (RE) through developing ontology which consists of some shared concepts. This research paper will assistance both of the RE suppliers and the customer to find each other semantically.

KAON’s text mining is a well-known tool that has been used. KAON defined as an open-source ontology management infrastructure aimed mainly for business applications. It includes a comprehensive and complete tool suite letting easy ontology creation and ontology management and provides a framework for building applications based on ontology. A crucial issue of KAON tool is that it is both scalable and efficient reasoning with ontologies [5]. KAON tool comprises of a few different modules providing a broad bandwidth of functionalities centered on creation, retrieval, storage, maintenance, and application of ontologies. KAON was in the past decade and still being further developed in a joint effort mainly by members of the Institute AIFB at University of Karlsruhe and the FZI – Research Center for Information Technologies. In this research paper five popular renewable energy providers were chosen to be the source of our data. From these different sites, the main concepts were extracted and then reduced into a smaller group of concepts which share common semantics.

This structure of the paper is as follows: section two introduces the background and motivation for developing semantic concepts. Section three recognizes the main steps of our first experiment and techniques that includes data collection, results analysis, and validation. Section four proposes a new technique data collection and evidences regarding them. Finally, section five concludes this research paper.

2. Background and Motivation

In last decade, a lot of researchers have proposed a new trust models. From these models it was noticed that, semantic trust share was very low. Salam et al. [6] stated that trust is crucial and critical to the spread and success of e-commerce. As the transactions done on-line are all about trust, the consumers of these transactions required to share sensitive personal and financial information with the service or product suppliers on the other side. Many other researchers also have worked on the issue of trust of the web sites as it is becoming more critical issue in doing financial transactions [21 - 23].

Certain trust standards and measures for e-commerce have developed and validated by McKnight et al. [7]; they have categorized them into four high-level classes: disposition to trust, institution-based trust, trusting beliefs, and trusting intentions. The psychometric properties of the measures are demonstrated through the use of a hypothetical legal advice website. The achieved results show that trust is surely a multidimensional concept. On other hand, Xiaochun et al. [8] have showed that one of the main reasons for many consumers to be worried from doing processes and online transactions is the issue that their financial information and personal information are being shared by other people on other other side of the net. Since it is not enough to know and make sure that the site you are communicating with offers encrypted communications through the closed padlock sign or the use of HTTPS Protocol or might has an address that matches the address on the certificate it has. What worries the consumer is the credibility of the site itself. Xiaochun et al. [8] have developed a suite of mechanisms of trust to reduce consumers’ suspiciousness and risk while they shop online. Matthew et al. [9] have proposed a theoretical trust model for Business-to-Commerce. The proposed model is based on four different categories: trustworthiness of the internet (the shopping medium), infrastructure factors (such as security, fraud, certificates from authorities), trustworthiness of the commercial, beside other different factors including demographic variables, company size, previous experience, etc. The results that shown by their work proved that the commercial integrity is a one of the major factor in that concerns consumer trust. Samia Nefti et al. [10] have proposed a new model that helps in choosing the needed and required information that the consumer would expect on a commercial website in order to complete a transaction in trustable issue. The model is based on fuzzy logic to estimate the trust of e-commerce sites; they showed and argued that the fuzzy logic is suitable and convenience to measure trust since it takes into account the uncertainties of the data as in human relationships.

The semantic web is very crucial and it has added standards and values to the WWW by giving a significant value to the information and data found on the net. This is achieved by adding a metadata (machine understandable content) to the web. Metadata is based on a domain ontology (a domain's conceptualization agreed upon by a community) [11]. From the achieved results in this research paper, it was included that some recent works in the last few years have examined factors that affect trust in different types of web sites, including e-commerce sites [12, 13]. Other work has looked at the site’s credibility in ways that are considered to be too limited to draw robust conclusions [14]. Many researchers have proposed means of evaluating the quality of web information, while web site consultants have proposed different ways to make a web site more credible and trustable [15, 16, 24].

Following from the above information, we still think that still there is shortage in a semantic trust model that could be applied in order to develop and enhance the trust between the customer and the seller.

3. The Experiment

In this paper, an experiment has been conducted to extract concepts for RE providers through text-mining tools in the domain of online RE providers. The following steps summarize the work that has been done:

1. Collect raw data for our domain and extract all concepts.
2. Refine the results to a smaller number of concepts
3. Validate the coverage of those concepts within popular providers.
4. Check those concepts against other non-popular ones.

In the following sections, we will address each one of the above steps.

A. Collecting **raw data and extracting top concepts**

List of five renewable energy companies (ordered by their stock exchange listed in Wikipedia1), were chosen to be the source of our data from. These providers are highly popular and have a very good reputation based on customer feedback, product selection, website techniques, shipping, payment options, customer support, and return policy. For more information and other reasons to our selection, please refer to [17, 18]. A2Z Group is an engineering, procurement and construction company, headquartered in Gurgaon, India, in Haryana state. Its inception was in 2002, and as of 2011 had more than 31,000 staff members [2]. The company has clients in the retail, government, infra-services and renewable energy sectors. Starting with A2Z Group, a sample of 35 pages was exported from the site; KAON's Text-To-Onto tool has been used to extract the concepts out of these pages, the size of the text pages exported was around 5.2 MB. Prior to applying the tool, the data was filtered by striping the HTML tags as well as the JavaScript codes from the files. This resulted in a set of files that are equal to about 670 KB in size. The tool generated more than 4600 concepts. For all these concepts, the number of times each concept was repeated, N, was also recorded.

B. Refining **the Results**

A second-level of filtering was implemented to refine these concepts by removing the duplicates as well as stop words, such as (at, the, of, etc). The number of concepts went down to 403 concepts. A third-level of filtering was made to remove unnecessary words/letters such as acronyms and vowels. This resulted in 144 concepts. Table 1 shows a sample of the list of 144 concepts concluded. For the complete list, please refer to [19].

TABLE 1: SAMPLE OF THE 128 CONCEPTS

Sample of Concepts				
wind water	climate	electricity genre	ethanol	Security
World	construct	energy	farm	Station
Year	core	energy efficient	company	phenomenon
agency	cost	energy secure	capacity	Form
Article	demand	energy source	application	Fossil
biomass	desert	energy supply	efficiency	Fuel
Build	develop	energy world	source	Future
carbon	earth	environ	statistics	generation

C. **Validating** the **coverage of those concepts**

In order to validate the coverage of the final list of concepts, the other popular providers were used. For each source, a set of 20 pages were selected, filtered and refined the same way we did with the pages from A2Z Group. A customized algorithm was developed to check the coverage of the 144 concepts extracted in section B above among those generated from the other source of data. Two pieces of information per concept were generated. These are: is the concept found or not, and what is the number of occurrences of the concept in question. We found that 59 out of the 144 concepts are common, i.e. shared among all five providers. The code of the algorithm is as shown in Table 2 below:

TABLE 2 ALGORITHM CODE TO CHECK THE COVERAGE

```
Dim found, occ, pos As Integer
Dim l, r As String
For i = 1 To 144
    currword = Trim(Sheet1.Cells(i, 1))
    Open "C:\Users\eyad.saleh\Documents\MT\AbeRE\all.txt" For Input As #1
    found = 0: occ = 0
    Do Until EOF(1)
        Input #1, s
        pos = InStr(1, s, currword, vbTextCompare)
        If pos <> 0 Then
            If pos = 1 And Len(s) = Len(currword) Then
                occ = occ + 1
            ElseIf pos <> 1 Then
                l = Mid(s, pos - 1, 1)
                r = Mid(s, pos + Len(currword), 1)
                If l = " " Or l = ">" Or l = "<" Or l = "-" Or l = ":" Or l = "/" Or l = "=" Then
                    If r = " " Or r = ">" Or r = "<" Or r = "-" Or r = ":" Or r = "/" Or r = "=" Then
                        occ = occ + 1
                    End If
                End If
            End If

    Loop
    Close #1
    Sheet1.Cells(i, 2) = occ
Next
```

We have noticed that the number of occurrences of the 59 concepts is not consistent among the five providers. For instance, the number of occurrences, N, for the concept "energy" in A2Z Group is 546, while in Provider 2, Provider 3, and Provider 4 it is 6, 8, and 3, respectively. Therefore, we have removed such inconsistent concepts after, of course, normalizing the number of occurrences per concept. The normalization process, that is changing N to f (frequency) is mainly to reduce the effect of the differences in the number of pages per site exported. Therefore, the concepts that are kept should satisfy the following formula:

$$\frac{A_i - f_{i,j}}{A_i} < 60\% \qquad (1)$$

Where fi,j is the frequency of the ith common concept in the j RE providers and Ai is the Average frequency of the ith common concept over the five RE providers. In turn, fi,j is defined by the following formula:

$$f_{i,j} = \frac{N_{i,j}}{\sum_{n=1}^{M} N_{n,j}} \qquad (2)$$

Where Ni,j is the number of occurrences of the ith concept in the jth RE providers, and M is the total number of concepts. In this case, M is set to 48.

TABLE 3. N VALUES FOR A SAMPLE OF CONSISTENT CONCEPTS

Concept	Wind water	climate	electricity genre	ethanol
Provider 1	57	56	44	69
Provider 2	57	64	145	47
Provider 3	135	379	167	105
Provider 4	44	29	12	20
Provider 5	31	41	23	40

Table 3 lists the N values for a sample of concepts that are considered consistent. This process resulted in a final set of 45 concepts to produce our ontology. These concepts are listed in Table 4.

TABLE 4. ONTOLOGY CONCEPTS IN THE FIELD

Final List of Ontology Concepts				
wind water	climate	electricity genre	ethanol	security
world	construct	energy	farm	station
year	core	energy efficient	company	phenomenon
agency	cost	energy secure	capacity	form
article	demand	energy source	application	fossil
biomass	desert	energy supply	efficiency	fuel
build	develop	energy world	source	future

Figure 1 illustrates the average frequency (among all four sites) histogram (in percentage) for all 45 concepts representing our Ontology. It is worth to note that the frequency is measured while M, in Equation 2, is set to 45, while the real value should be when M is equal to 144 or more, as found in section B. However, for simplicity we set it to 45.

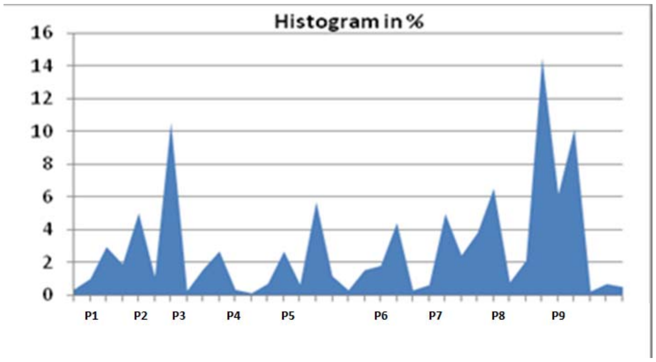


Fig. 1 Average Frequency Histogram.

A. *Testing the concepts*

We have chosen eight providers for our experiment. We added a ninth one to the list that was actually hosted under a real domain name specifically to perform this experiment. The site was created by fourth year students in the faculty of IT through the course of their graduation project. The students scored a high grade on their project due to its friendly design and were among those who competed for the best graduation project of the year. When the students were asked to upload their site to epicforsale.com, they were asked to connect it fully with a payment gateway (they chose PayPal) and populate their database with a good number of products at prices 10% lower than A2Z Group.

Three experts (RE specialists) as well as two regular users who are considered as “below average” as far as their internet usage is concerned were chosen to judge the trust of those nine websites. None of the users did actually commit any transaction. The feedback from the experts and the regular users are listed below in Table 5. The first column lists the URLs of the websites being tested. The second to fourth columns show the percent of trust for each website we received from expert 1 to 3. The fifth column shows the average of trust among the experts. The sixth and seventh columns show the percentages of trust we received from regular user one and two, respectively. Finally, the eighth column shows the average trust between both regular users. We could have increased the number of Experts to get a better average, but the values are not the focus of this experiment if compared to the results generated upon applying our Ontology.

TABLE 5. EXPERTS JUDGMENTS

URL	E1	E2	E3	Avg Expert	R1	R2	Avg Reg
Provider 1 URL	85 %	90 %	85%	87%	80 %	90 %	85%
Provider 2 URL	80 %	85 %	80%	82%	70 %	80 %	75%
Provider 3 URL	80 %	80 %	85%	82%	80 %	85 %	83%
Provider 4 URL	80 %	80 %	85%	82%	85 %	85 %	85%
Provider 5 URL	70 %	75 %	80%	75%	80 %	80 %	80%
Provider 6 URL	75 %	80 %	80%	78%	75 %	80 %	78%
Provider 7 URL	65 %	70 %	70%	68%	70 %	80 %	75%
Provider 8 URL	55 %	60 %	70%	62%	70 %	85 %	78%
Provider 9 URL	10 %	15 %	20%	15%	70 %	75 %	73%

Once we received the feedback, we applied the final set of the 45 concepts (refer to section 3.3) on all nine sites. The coverage of each concept was calculated and compared; that is, the frequency of each concept in these nine websites was compared against the average frequency of the four popular RE provider sites (Figure 1). In other words, the histograms are compared together. The Euclidean distance between two histograms x and y is computed as follows:

$$D(x,y)=\sqrt{(\sum(x(n)-y(n)^2))} \qquad \textbf{(3)}$$

Where x(i) and y(i) represent the frequencies of the ith concept. Figure 2 shows the D values for each of the nine sites.

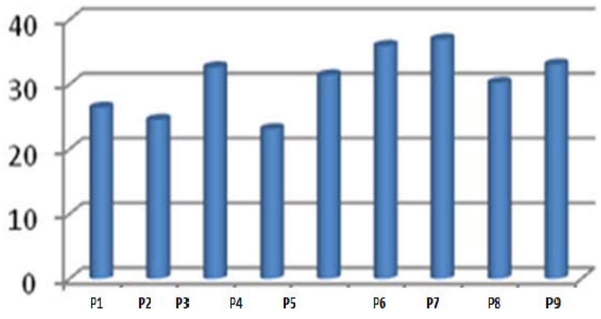


Fig. 2 D values for the nine tested web sites.

From Figure 2, one cannot find a proper threshold to compare the value of D with, which would indicate whether the site is trustworthy or not upon crossing. This is because the histograms we are constructing are not the accurate histograms that we would expect. This is because, as mentioned earlier, M in equation 2 is set to 45 for simplicity. This value might be reasonable to A2Z group from which we drew our 45 concepts from, but less reasonable to the other four popular RE providers, and may be far from being reasonable for the others. This is because the calculated histogram value x(i) using M=45 is close but not exact in value to that when using M=144 as expressed in Equation (4).

$$x(i)=N_i/\sum N_{n\ (n=1\ ..\ 45)} \neq N_i/\sum N_{n\ (n=1\ ..\ 144)} \qquad (4)$$

Having that in mind and replacing D with R to represent the summation of the minimum frequency values for each concept i between two histograms x and y gives us the following equation:

$$R=\sum \mathrm{Min}(x(n),y(n)) \qquad \textbf{(5)}$$

Figure 3 shows the R values against all nine sites. Choosing 35% as the threshold, we would grant trust to all but the last one. This confirms both the Experts and the regular users’ judgements for the first eight sites; while in the last one, the results have confirmed the experts’ but not the regular users' judgements on that site.

Thus, our ontology for RE provider’s domain has helped in distinguishing between trustworthy and untrustworthy sites. The results were on the same page as the experts’ judgments, yet not deceived by the nice or friendly design of the ninth site just like it deceived the regular user. It simply means that our ontology, to some extent, is useful and could be used as an assistant to check the trustworthiness of RE providers.

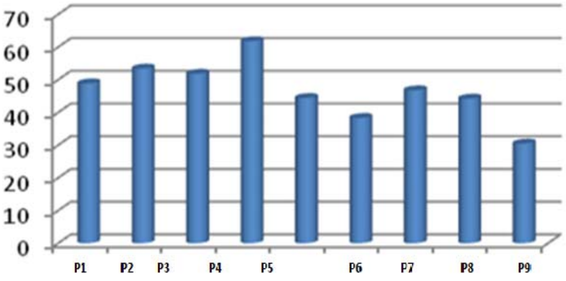


Fig. 3 R values for the nine tested web sites

4. New Efficient Technique

Our works show that with a threshold 35% we can reach to the same results of the experts. On the other hand, the threshold can’t give us how much this provider is close to the common concepts. For that we will barrow the definition of coverage from Kayed [25] and deploy it here. Let’s say that the resulted 45 concepts for the four providers is our ontology. Therefore, a coverage measure can be used to evaluate our proposal. Our aim is to make sure that our proposed ontology-concepts are “good” and can be used later. We need to check whether the concepts of new providers are well covered by our proposed ontology-concepts or not.

Let the coverage measure, Cov(D,O) represent the number of concepts (not words) in a description D that are covered by the ontology O. We distinguish between the terms words and concepts. The concept term refers to the word itself and all words that have any relationship with the word such as synonyms. Furthermore, in the process of matching and counting, we are using the concepts rather than just the words, i.e. the words and their meanings, using WordNet [26] ontology.

This measure has been used to evaluate the providers’ concepts. In following, the measure will be further elaborated. Later, we will apply it to check the closeness of the provider concepts with regards to its domain concepts.

Definition: Let O be an ontology, D be description for a collection of concepts, w_1, w_2, …, w_m be words in the collection, and N be the number of concepts in the description D. Let also the functions: mi, F, and Cov(D, O) be defined as follows:

$$\mathbf{m_i} = \begin{cases} 1 & w_i \in O \\ 0 & \text{Otherwise} \end{cases}$$

$$\mathbf{F} = \sum_{i=1}^{m} m_i$$

$$\mathbf{Cov\ (D,\ O)} = \frac{F}{N}$$

The Cov(D, O) will read as the coverage of an ontology O for a description D. Cov(D, O) measures how many concepts in a description D that belong to an ontology O. It calculates the number of words in a description D that matches concepts (i.e. words or their meanings) in the provided ontology O.

The number of these matches will be stored in the variable F. The Cov(D, O) will be normalized by dividing F on the number of concepts in the description D, that is N. This measure computes the number of matches of the words with tier meanings in the description D to the number of concepts in the description D. If Cov(D, O) is close to 1 this means that the description D is very close to be covered by the ontology O. Cov(D, O) gives how much the description D is covered by the concepts of the ontology O.

The coverage Cov(D, O) has been computed for all nine providers. The ontology O will be the common concepts for the four providers. For each provide, we collect the description concepts and we used WordNet to get their meanings. Then we compute how far each provider is from the common concepts. Table 6 illustrates that. The absolute error has been computed and gives an average of 5%. This means that the coverage measure is close to the expert with an error 5 % on average.

TABLE 6. COVERAGE MEASURE.

Providers	Avg. Expert	Cov(D, O)	Abs_error
P1	0.87	0.91	0.04
P2	0.82	0.87	0.05
P3	0.82	0.92	0.1
P4	0.82	0.88	0.06
P5	0.75	0.77	0.02
P6	0.78	0.8	0.02
P7	0.68	0.63	0.05
P8	0.62	0.65	0.03
P9	0.15	0.25	0.1
		Avg_error	0.05

5. Conclusion and Future Work

One of the main important and critical issues to the success of both e-commerce and e-business applications is trust. All over the world numerous number of people are using the internet to do their financial transactions online; anyway these transactions can't be 100% sure of the credibility, reliability, or trustworthiness of the commercial they are using especially when it is new. As contribution for this issue, our research has concentrated on the domain of Renewable Energy providers. We argued and claimed that using semantic and ontologies to measure the trust of online RE providers is crucial and useful. As KAON's tools used to extract concepts from the different data sources, in this research paper a final set of 45 concepts were chosen and checked against nine providers. The results which were achieved were compared against three experts' judgments as well as two non-expert users' judgments. Finally, the experiments and the new technique in sections 4 and 5 have showed that the results of our approach matched that of the experts' and was able to pull out a test online RE providers from a pool of real ones. The test site is represented by a group of undergraduate students' graduation project which does not reflect a real online business.

In a future work, a formal representation of the ontology will be developed; relationships among the concepts will be build, evaluate and enhance the ontology by using other approaches. Furthermore, computerize all the steps mentioned in our work through building a friendly tool. This tool will help by the online users to assist them in examining the credibility of the different sites they might do transactions on.

References

[1] Fogg, B.J., Marable Leslie, Julianne Stanford and Ellen Tauber, “How Do People Evaluate a Web Site’s Credibility? Results from Large Study”, (**2002**).

[2] Fogg, B.J. & Tseng, H. “The Elements of Computer Credibility”. Proceedings of the CHI99 Conference on Human Factors and Computing Systems**,** *ACM Press***, (1999),** pp. 80-87.

[3] Morkes, J. and Nielsen, J., Concise, “SCANNABLE and Objective: How to Write for the Web”, *See www.useit.com/papers/webwriting/writing.html*, (**1997**).

[4] Fogg, B.J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., et al.. “What Makes Web Sites Credible? A Report on a Large Quantitative Study”. *Proceeding of the CHI01 Conference of the ACM/SIGCHI*, Seattle, WA: ACM Press, **(2001**), pp. 61-68.

[5] KAON’s homepage*, http://www.kaon.semanticweb.org*

[6] A. F. Salam et el, “Trust in e-commerce”, *Communications of the ACM,* Volume 48, Issue 2 (**February 2005**).

[7] D. Harrison McKnight, Vivek Choudhury, Charles Kacmar, “Developing and Validating Trust Measures for e-Commerce: An Integrative Typology”. **(2002)**. pp. 334-359.

[8] Xiaochun Zeng, Jiaoyan Zeng, Qiang Guo. “Research of trust on B2C electronic commerce”. ICEC' (**2005)**. pp.221~225.

[9] Matthew K. O. Lee and Efraim turban, “A Trust Model for Consumer Internet Shopping”, *International Journal of Electronic Commerce*, Volume 6, Issue 1, Number 1/Fall **(2001)**.

[10] Samia Nefti, Farid Meziane, Khairudin Kasiran, “A Fuzzy Trust Model for E-Commerce”, *Seventh IEEE International Conference on E-Commerce Technology*, **(2005)**.

[11] Daniel Oberle et el, “Supporting application development in the semantic web, *Transactions on Internet Technology* (TOIT)”, Volume 5 Issue 2, May (**2005)**.

[12] Cheskin Research, “Trust in the Wired Americas”, *See www.cheskin.com/think/studies/trust2.html*. (**2000).**

[13] Cheskin Research, S.A.S., “eCommerce Trust Study”. *See www.studioarchetype.com/headlines/etrust_frameset.html.* (**1999).**

[14] Critchfield, R. “Credibility and Web Site Design”, See www.warner.edu/critchfield/hci/critchfield.html, (**1998)**.

[15] Wilkinson, G.L., Bennett, L.T., & Oliver, K.M. “Evaluating the Quality of Internet Information Sources”. *See http://itech1.coe.uga.edu/faculty/gwilkinson/criteria.html,* (**1997).**

[16] Nielsen, J., “Trust or Bust: Communicating Trustworthiness in Web Design”, *See www.useit.com/alertbox/990307.html,* **(1999)**.

[17] www.top10listguide.com

[18] http://en.wikipedia.org/wiki/List_of_renewable_energy_companies_by_stock_exchange#List_of_publicly_traded_renewable_energy_companies).

[19] Drkayed@ymail.com

[20] http://en.wikipedia.org/wiki/List_of_renewable_energy_companies_by_stock_exchange#List_of_publicly_traded_renewable_energy_companies, (**2013).**

[21] Shneiderman, B. Designing trust into online experiences. Comm. *ACM* 43(12), . (**2000),** pp. 57–59.

[22] Rousseau, D. M., S. B. Sitkin, R. S. Burt, C. Camerer. Not so different after all: A cross-discipline view of trust. Acad. Management Rev. (**1998)**, 23(3) 393–404.

[23] Stewart, K. J. Transference as a means of building trust inWorld Wide Web sites. P. De and J. I. DeGross, eds. *Proc. 20th Internat. Conf. Inform. Systems,* December, Charlotte, NC, (**1999)**, pp. 459–464.

[24] Egger, F. N., Affective Design of E-Commerce User Interfaces: How to maximise perceived trustworthiness. *Conference on Affective Human Factors Design*. Singapore, June 27-29, (**2001)**, pp. 317-324.

[25] Ahmad Kayed, Nael Hirzallah; Mohammad Alharibat, Ontological-Concepts-Coverage Measure for Software-Component Descriptions, *European Journal of Scientific research, EJSR*(Volume 101 Issue 1, **(2013)**, pp77-99.

[26] http://wordnet.princeton.edu/, Sep., 201.

Using Graphs to Visualize Overlapping Sets

Abdalmunam Abdalla[1], Mehmet Koyuturk[2], Abdelsalam M. Maatuk[3], Alfaroq O. Mohammed[4]

[1,4]Department of Computer Science, Omer AL-Mukhtar University, Libya
[3]Faculty of Information Technology, Benghazi University, Libya
[2]Faculty of Engineering, Case Western Reserve University, USA

Abstract: The large volume of data available in many domains and the need to analyze the data to extract useful information from it has lead to the need of visualization techniques to get information about the data at a glance. Visual inspection is useful in providing fast and abstract information about datasets to guide the researchers in choosing the suitable approach to process the data. Recently, there have been notable advances in graph visualization; however, visualizing sets still needs more attention. In this paper a method is proposed to visualize overlapping sets so that the underlying hierarchy and relations of the sets can be easily understood by visual inspection. This approach utilizes the graph representation of the sets to aid the drawing process. Using the spectral decomposition of the graph derived from the sets, we developed algorithms to compute the best coordinates for the items of the sets and plot them on the Euclidean plane. The method has been tested on both real and synthetic datasets to investigate its performance.

Keywords: Sets, sets visualization, overlapping sets, visualization, sets drawing.

1. Introduction

Visualization is a mean of representing data, so that data can be explored and understood by visual inspection. Visualization utilizes the human visual ability to allow users to understand visualized objects and their underlying relations. Visual representation of data is useful in providing abstract information about the data at once. Visualization is increasingly applied in many applications, including software engineering [1], imaging [2], digital libraries [3] and others.

In this paper, we devise algorithms to visualize overlapping sets; hence the users (e.g. researchers) can easily and quickly have information about the underlying relationship between the sets. The visualization of sets can help to classify them and make decisions for further analysis. For example, we can see which subsets of the sets are heavily overlapping, so that we can study them together because they share a lot of information. In contrast, disjoint sets can be analyzed separately.

This problem is closely related to graph visualization and multi-dimensional scaling (MDS). However, this problem is fundamentally different in that we are also interested in visualizing the sets in addition to their items. In MDS, the distances between items are given and a low-dimensional mapping of the items is sought. In graph visualization, the pairwise relationships between items are given and a two-dimensional mapping of the items and the pairwise relationships is sought. On the other hand, this study intends to obtain a two-dimensional mapping of the items in such a way that we can visualize which items belong to which set. This requires the optimization of the mapping of the items on the two-dimensional space, as well as the optimization of the representation of the sets. To verify the correctness of the algorithms, experiments have been conducted on both real and synthesis datasets, which showed a positive indication on the efficiency of the proposed method.

The rest of the paper is organized as follows. Section 2 provides an overview of related work. The problem is defined formally in Section 3. Section 4 presents our spectral approach for sets visualization based on the Laplacian of co-membership and bipartite graphs generated from the sets. The experimental results, performance and evaluation metrics used are discussed in Section 5. Section 6 concludes the paper.

2. Related Work

A great amount of research has been recently done on information visualization. In this section we discuss the work related to set visualization, including visualizing sets, graph visualization, and multidimensional scaling.

Euler diagrams can be used to represent the relationships between sets. Many algorithms have been developed for drawing Euler diagrams. Flower and Howse [4] outlined well-formedness conditions on drawn diagrams and presented an algorithm to decide whether or not an abstract diagram is drawable under those conditions. If a diagram is diagnosed as drawable, then a drawing is produced. Later work [5] aimed to enhance the layout of an already drawn Euler diagram, using a hill-climbing based optimization approach in combination with a range of layout metrics, to assess the quality of the drawing. An Euler diagram drawing tool [6] designed, which embeds some small diagrams which can be drawn with a limited subset of shapes. Alsallakh et al. [24] proposed a technique for finding and analyzing different kinds of overlaps between sets using frequency-based representations. An algorithm based on shortest-path graphs to depict set membership of items on a map has been presented by Meulemans et al. [25]. Lex et al. [26] introduced a novel visualization technique for the quantitative analysis of sets, their intersections, and aggregates of intersections.

Simonetto et al. [7] developed another algorithm to generate Euler-like diagrams. This algorithm differs from others in that it has no un-drawable instances of the input. Such algorithms were mainly designed to draw Euler diagrams for a very small number of sets with small sizes. To date, many approaches to graph drawing have been developed [8, 9]. Many kinds of graph-drawing problems exist, including drawing directed graphs, drawing planar graphs, and others.

An existing tool, daVinci [10] is used as a user interface for graph layout in many applications. The tool has some useful features such as scaling operations, abstraction, and fine-tuning of the layout. Clemencon et al. [11] described a methodology for graph visualization based on hierarchical maximal modularity clustering. Luo et al. [12] proposed an ambiguity-free edge-bundling technique to improve the visualization of very dense graphs. The method is useful in producing clear visualization of complex graphs that is caused by the density of the edges.

Many spectral approaches for graph visualization exist, that use the eigenvectors of the graph matrix to produce a mapping of the graph vertices to the Euclidian space [13, 14]. Koren [15] developed an algorithm that uses the eigenvectors of the Laplacian to visualize graphs. They utilize the eigenvectors of the Laplacian of the graph to compute the best Euclidian coordinates of the items of the sets.

MDS methods are used to decrease the dimensionality of data while preserving as much information as possible about these data. Many methods [16,17,18] have been developed for solving the nonlinear dimensionality reduction problem.

Leeuw and Mair [19] proposed a solution to the multidimensional scaling problem by means of the majorization algorithm. This method is intended to minimize the stress and functions to majorize stress were elaborated. Agarwal et al. [20] introduced an iterative local improvement method for solving many variants of multidimensional scaling problems. The algorithm starts by choosing a point and moving it so that the cost function is locally optimal and repeats the procedure until convergence is achieved. Chen and Buja [21] proposed a local multidimensional scaling method that constructs a global embedding from local information. The method localizes versions of MDS stress functions by using force paradigm and a tuning parameter.

In contrast, we are considering the problem of visualizing a large number (i.e., tens to hundreds) of larger (e.g., tens of items) sets and we also consider undrawable cases. In order to handle undrawable instances, we might allow errors. We are used a graph-drawing algorithm to visualize overlapping sets. The methods we develop are based on spectral decomposition of the graph representation of sets. In addition to the two-dimensional mapping of the items on the Euclidian plane, we require the optimization of the representation of the sets, so that we can visualize the items and sets together.

3. Problem Definition

Formally, we have n items grouped into m sets where $Si = \{Ij\}$, $1 \leq i \leq m, 1 \leq j \leq n$. We desire to plot these items on the Euclidean plane and group them in a way that makes it to see the relationship between these items easily. Without loss of generality, it is required that the co-membership graph produced by the sets to be visualized is connected. If the graph is not connected, we pre-process the data and provide multiple connected graphs and visualize them separately. This is natural because we are interested in visualizing the overlap between sets, thus we do not have to visualize disjoint sets in the same drawing.

Example 1: Consider the following group of sets and items. We refer to the *i-th* set as S_i and the *j-th* item as I_j.

$$S_1 = \{I_1, I_2, I_3\}$$

$$S_2 = \{I_1, I_2, I_5\}$$

$$S_3 = \{I_3, I_4\}$$

These sets need to be visualized, so that each item is represented by a point, and each group of items belong to a set is bounded by a circle that represents that set.

4. Proposed Algorithms

This section presents our method for visualizing overlapping sets. The algorithm regards the sets as a graph and visualizes them using the Laplacian of the graph. Namely, we use the eigenvectors to compute the optimal position for the graph nodes (items) as well as the circles that represent the sets in the two-dimensional plane. Drawing the sets by randomly assigning values to the x and y coordinates of each item in the sets does not provide useful information about the sets and their relations. Even with a very small number of items the drawing does not represent the dataset correctly.

In this study, we have proposed several algorithms based on the spectral decomposition of the graph representation of sets. We first start with the algorithm, which is based on a co-membership graph, and then we perform several improvements to this algorithm.

4.1 Algorithm based on co-membership graph representation of sets (CMG)

Since many of the existing algorithms aim to visualize graphs, we first represent the sets to be visualized as a graph. For this purpose, we construct the co-membership graph of the sets. Figure 1 shows the co-membership graph of Example 1 given in Section 3.

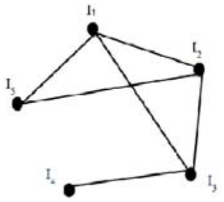


Figure 1. Co-membership graph of the three sets in Example 1

To construct the graph, we first construct the membership matrix M of the group of items. M is an $m \times n$ matrix, where m is the number of sets and n is the number of items. If item I_j is contained by set S_i, the entry M_{ij} is set to one, otherwise, it is set to zero. For example, the membership matrix for the sample instance given above is the following:

$$M = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \qquad (1)$$

From the membership matrix, we can construct the adjacency matrix U of the co-membership graph G. U is an $n \times n$ matrix, where each entry U_{ij} is set to one if both items i and j appear together in at least one set, or it is set to zero if the two items do not share a set. Besides, all the diagonal entries U_{ii} are set to zero. The adjacency matrix of Example 1 is as follows:

$$U = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix} \qquad (2)$$

This matrix is called the unweighted adjacency matrix, and it can be computed directly from M using the following equation:

$$U = M^T \otimes M \qquad (3)$$

Here, the operator $\otimes$ specifies a modified matrix multiplication, in which multiplication is replaced by logical “AND” and the addition is replaced by logical “OR”. If both items i and j appear in the same set more than once (appear together in more than one set), it might be useful to include that information in the co-membership matrix, since the items that appear in many sets together should be located close to each other in the two-dimensional plane. Therefore, we define a weighted co-membership matrix W, in which the entry W_{ij} will be equal to the number of sets that contain the two elements together. Once more, all the diagonal entries W_{ii} are set to zero. We can easily compute W directly from M by the following equation:

$$W = M^T \times M \qquad (4)$$

The weighted adjacency matrix for Example 1 is the following:

$$W = \begin{bmatrix} 0 & 2 & 1 & 0 & 1 \\ 2 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix} \qquad (5)$$

We then construct the Laplacian matrix L of the co-membership graph. The Laplacian matrix is an $n \times n$ matrix where $L = D - A$ and D is the degree matrix, which is an $n \times n$ diagonal matrix where D_{ii} = deg(i). The Laplacian L can be computed from the unweighted adjacency matrix U as follows:

$$L_{ij}^{(U)} = \begin{cases} \deg(i) & i = j \\ -U_{ij} & i \neq j \end{cases} \quad i,j = 1,2,\dots,n \qquad (6)$$

Where deg(i) in this equation denotes the degree of item i (the number of items that appear with item i in at least one set). Similarly, the Laplacian can be computed from the weighted adjacency matrix W as follow:

$$L_{ij}^{(W)} = \begin{cases} \mathrm{wdeg}(i) & i = j \\ -W_{ij} & i \neq j \end{cases} \quad i,j = 1,2,\dots,n \qquad (7)$$

Where wdeg(i) in this equation denotes the weighted degree of item i (the sum of the weights of the edges incident to i).

Using the Laplacian is very useful in that we convert the problem to an optimization problem. Consider the problem of mapping the nodes of a graph onto one-dimensional Euclidian space such that the nodes, which are connected with heavier edges are closer to each other in the space. This problem can be formulated as follows:

$$\min_x E(x) \overset{\text{def}}{=} \sum_{\langle i,j \rangle \in E} w_{ij}(x(i) - x(j))^2 \qquad (8)$$

The right-hand-side of the above equation can be written in matrix form as follows:

$$x^T L x = \sum_{i<j} w_{ij}(x_i - x_j)^2 \qquad (9)$$

Therefore, since L is a positive semi-definite matrix, the eigenvector corresponding to the second smallest eigenvalue of the Laplacian provides the optimal solution to this problem. Similarly, the optimal solution to the problem of mapping graph nodes can be computed into two-dimensional Euclidian space by taking the eigenvectors that correspond to the second and third smallest eigenvalues of the Laplacian. In Example 1, if we use the unweighted adjacency matrix for the co-membership graph, the Laplacian will be as follows:

$$L^U = \begin{bmatrix} 3 & -1 & -1 & 0 & -1 \\ -1 & 3 & -1 & 0 & -1 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ -1 & -1 & 0 & 0 & 2 \end{bmatrix} \qquad (10)$$

For the weighted adjacency matrix, the Laplacian is the following:

$$L^W = \begin{bmatrix} 4 & -2 & -1 & 0 & -1 \\ -2 & 4 & -1 & 0 & -1 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ -1 & -1 & 0 & 0 & 2 \end{bmatrix}. \qquad (11)$$

Koren [15] showed that the eigenvectors of the Laplacian are very useful in providing a comprehensible visualization of graphs. Koren [15] also showed that using the degree-normalized eigenvectors of the Laplacian provided more natural visualization than using just the eigenvectors of the Laplacian. Normalization means that we consider the degree when we compute the eigenvectors from the Laplacian. Here we require that

$$x^T D x = 1 \text{ and } y^T D y = 1 \qquad (12)$$

Further details on deriving the eigenvectors from the Laplacian matrix can be found in [15]. In order to comprehensively investigate the use of co-membership graphs in visualizing sets, we use four variants of the co-membership graph G: unweighted, weighted, unweighted normalized and weighted normalized. The performance of the algorithm for each variant of the graph is described in Section 5.

After computing the optimal x and y coordinates for all the nodes (items) of the graph, we use circles to represent the sets. Each group of items belongs to one set is bounded by a circle. In order to draw these circles, we need two quantities: the center of each circle and the radii of the circles. For each circle (set), the average of the x coordinates of the items that belong to that set is computed and used as the x coordinate of the center of the circle representing that set. Similarly, the y coordinates for each circle (set) are computed. The radius of the circle is the distance between the center and the furthest item from the center in the set. Computing the radii in such a way guarantees that all the items in a set are bounded by the circle representing that set, allowing no false negatives, but allowing some items that are not in the set to be visualized as if they have been in the set.

4.2 Algorithm based on bipartite graph representation of sets (BPG)

In the CMG algorithm, there are some cases, in which to instances of different datasets produce the same co-membership graph. For instance, consider the following group of sets:

$$S_1 = \{I_1, I_2, I_3\},\ S_2 = \{I_1, I_5\},\ S_3 = \{I_2, I_5\},\ S_4 = \{I_3, I_4\}.$$

This instance produces the same co-membership graph as in Example 1 shown in Figure 1. In other words, different distribution of items to sets might give rise to the same co-membership graph, causing loss of information. To overcome this problem, we propose an algorithm that is based on the bipartite graph representation of the sets to be visualized, and then compute the Laplacian and eigenvectors. Consequently, the coordinates for the items and the centers of the circles representing the sets can be directly computed. In this algorithm, we use the following equation to compute the adjacency matrix of the bipartite graph:

$$A^T = \begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix} \qquad (13)$$

Where M is the membership matrix and the size of A is $m + n \times m + n$. Note that the zero in the upper left corner of the matrix is an $m \times m$ zero matrix and the zero in the lower right corner is $n \times n$ zero matrix. The adjacency matrix of the bipartite graph in our example is

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \qquad (14)$$

We do not need edge weights since each entry in the adjacency matrix corresponds to exactly one set-member relation. After we construct the adjacency matrix, the Laplacian and the eigenvectors can be computed. The Laplacian for this adjacency matrix becomes as:

$$L = \begin{bmatrix} 3 & 0 & 0 & -1 & -1 & -1 & 0 & 0 \\ 0 & 3 & 0 & -1 & -1 & 0 & 0 & -1 \\ 0 & 0 & 2 & 0 & 0 & -1 & -1 & 0 \\ -1 & -1 & 0 & 2 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 2 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \qquad (15)$$

After this step, we compute the degree-normalized eigenvectors of the Laplacian. Subsequently, we compute the centers of the circles directly from the resulting eigenvectors of the Laplacian. Namely, we use the first m elements of the eigenvector that corresponds to the second smallest eigenvalue as the x coordinates of the centers of the sets and the first m elements of the eigenvector that corresponds to the third smallest eigenvalue as the y coordinates for the set centers. The remaining elements from the second and third low eigenvectors are used to compute the x and y coordinates for the items respectively, as in the CMG algorithm. We compute the radii of the circles in a way that is similar to CMG; hence, BPG is also guaranteed to have no false negatives.

4.3 Improved bipartite graph based algorithm (IBPG)

When drawing the circles, there might be items that lie far away from the center of the circle that represents a set and most of the other items seem to be clustered around the center. The number of those items is usually small, and including them in the circle increase the radius of the circle, which subsequently increases the size of the circle allowing more false positives (items that appear to be in the circle but do not actually belong to the set represented by that circle). As a tradeoff, we try to exclude those items as an effort to keep the number of false positives low. BPG algorithm outperforms CMG, thus we build IBPG on top of BPG. After we compute the x and y coordinates for the items and the circles, for each circle we construct the table shown in Table 1.

Table 1. Computing best radius

R	Tp	Fp	Fn	t
$\mathbf{r_1}$	tp_1	fp_1	fn_1	t_1
$\mathbf{r_2}$	tp_2	fp_2	fn_2	t_2
⋮	⋮	⋮	⋮	⋮
$\mathbf{r_k}$	tp_k	fp_k	fn_k	t_k

Here, k is the size of the set, tp is the number of true positives (i.e., number of the items that belong to a set and are bounded by the circle representing that set), fp is the number of false positives (i.e., number of items that do not belong to a set and visualized as they were in that set), fn is the number of false negatives (i.e., number of items that belong to a set and visualized as they were not in the set), and r_i represents candidate radii. The way we generate candidate radii is as follows: we start with one item in the set and compute the distance between that item and the center of the circle, and compute all the *tp, fp,* and *fn* associated with that radius. Then, we add the next closest item to the center, in the circle and choose the candidate radius as the distance between the center of the circle and the farthest item so far. Similarly, we compute the entries of the table of the chosen radius, and so on, until we compute all the table entries for all candidate radii. Using this table, we compute the optimal radius of each circle that produces the best value in the following equation:

$$t(i) = \frac{tp(i) - fp(i)}{size(S)} \qquad (16)$$

Obviously, this equation is trying to maximize the number of true positives and minimize the number of false positives, and this is what our algorithm intends to do, (we want to include the items that belong to the set in the circle and exclude the items that do not belong to the set from the circle). There are some cases, in which various radii have the same t value. In such cases, we choose the radius that minimizes the number of false negatives. In other words, we pick the radius that includes more items in the circle without increasing the false positive rate.

4.4 Improving visualization

There are some cases, in which some items might be plotted over each other, and this is natural because if two items occur in the same set of sets, they are mapped to the same point in the Euclidean space. We have isolated those items and plotted them separately using a discretization process, so that they can be easily identified on the drawing. Precisely, we sort the x and y coordinates of the items and map them to integers from 1 to n in increasing order. Subsequently, we plot these integer vectors that are derived from the real-valued vectors.

5. Experimental Study

This section describes the performance evaluation of the method presented in this paper. In order to measure the performance of the proposed algorithms, two indicators are used. The first indicator is the total number of errors produced by the algorithm. We compute the membership of the items in the sets indicated by the visualization and compare them with the actual membership of the items in the sets. As a result, we can apprise which algorithm represents the true membership of the items more accurately. The second indicator is the total area of the circles we use to represent sets. A smaller area means that the algorithm utilizes the plot area more efficiently. In addition, we also focus on the aesthetic aspect of the drawing produced by the methods, as the objective of these algorithms is to produce a drawing that is easy to understand.

To evaluate the performance of the algorithms, we use three classes of randomly generated datasets. We consider three factors when we generate the data, the number of items, the number of sets, and the average set cardinality. For each dataset, we fix two factors at a time, and vary the other factor to see how that factor affects the performance. Besides, we have tested our algorithms on three real datasets. The first one is set of the prime factors of the numbers between 100 and 200. The second dataset is a subset of the publications list of Mehmet Koyuturk from Case Western Reserve University obtained from his web page [27]. The third dataset is a group of six biological annotation datasets consisting of genes and their functional annotation according to Gene Ontology (GO). Details on the last dataset can be found on [22]. Further details on the experimental setup and data sets can be found in [23,28].

Experiments of the CMG algorithm on synthetic datasets show an increase in the number of errors as the number of items increases because the plot area gets crowded with items. Suddenly, the number of errors goes down because as the number of items grows largely, the overlap diminishes. The number of errors grows steadily with the number of sets. This is expected because the plot area gets crowded as the number of sets increases. The number of errors increases fast as the average set size grows because the overlap increases, then the number of errors saturates because the likelihood of a false positive goes down since the likelihood that an item will be in a set goes up.

The results show that the normalization of Laplacian by node degrees outperforms the non-normalized version both in terms of number of errors and area of circles. Moreover, the weighted version performs better than the unweighted one. That is because when we are considering weights, we actually are considering the number of times that any two items appear in the same set together and this affects the relation between the items. For these reasons, we then compare the weighted normalized version of the CMG algorithm against the BPG algorithm. Comparing the results shows that the BPG algorithm performs better than the CMG algorithm. This is due to the way that the graph is computed from the sets, namely the bipartite graph. By using Equation 13 to compute the bipartite graph from the sets, we get a matrix that represents the set-member relations explicitly. Using that matrix, we also compute the centers of the circles directly based on the eigenvectors of Laplacian.

The experiments also demonstrate that the IBPG algorithm performs better than the BPG algorithm. However, the use of IBPG depends on whether or not false negatives are allowed. Drawing algorithms are not necessarily intended to provide complete, very accurate information about the data; rather they are designed to provide initial basic information about the data that help researchers to decide how to analyze this data. Some errors might occur, and further analysis would reveal more information about the data. The visualization algorithm can help us to decide on what subset of the data we study together when we know initial information about their relations. Yet, the user can choose between the BPG and the IBPG algorithms depending on whether or not they desire to allow missing data.

Figure 2 shows the results of applying the three algorithms: the CMG algorithm, the BPG algorithm, and the IBPG algorithm to a dataset comprised of a list of ten papers written by eighteen authors.

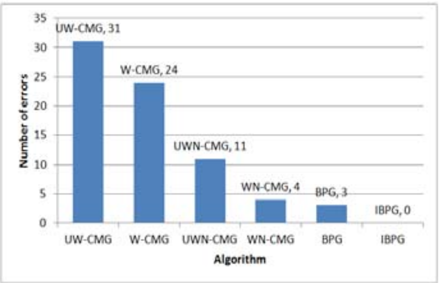


Figure 2. Performance of the three algorithms on the publication list of ten papers written by eighteen authors

Figure 3 shows the drawing of this data set obtained by applying the IBPG algorithm.

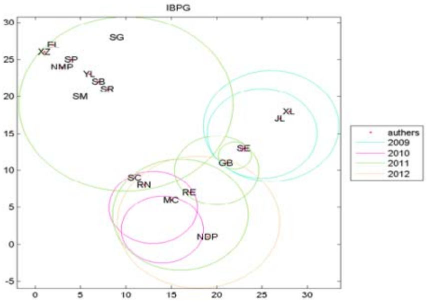


Figure 3. A drawing of the publication list obtained by the IBPG (18 authors and 10 papers)

From the figures above, we can see that the IBPG algorithm produces a visually comprehensive drawing. Figure 4 shows the performance of the three algorithms on the gene annotation datasets described previously. Yet again, we can see that IBPG algorithm performs better than the other algorithms, illustrating the benefits of using the set membership information directly as opposed to using co-membership graphs.

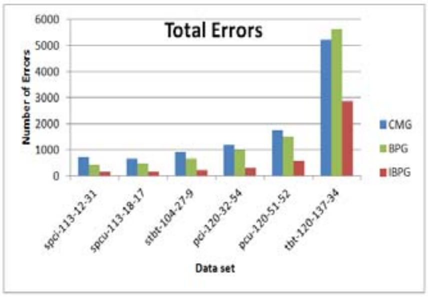


Figure 4. Performance of the three algorithms on the gene annotation datasets

We have enhanced the algorithms by adding extra aesthetic features. We are able to display only a group of circles that overlaps with a particular circle. This is useful because it reduces the number of objects on the drawing so we can get clearer results. Figure 5 shows which sets overlap with set number 2. In other words, it shows which papers share authors with paper number 2?

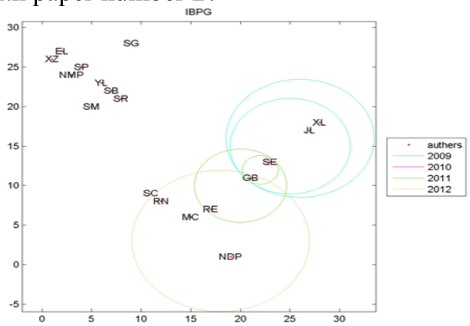


Figure 5. Papers that overlap with paper number 2 in the publication list

Another feature is to pick up any arbitrary author and see which authors share papers with that author (co-authorship). Figure 6 shows the co-authorship of G. Bebek.

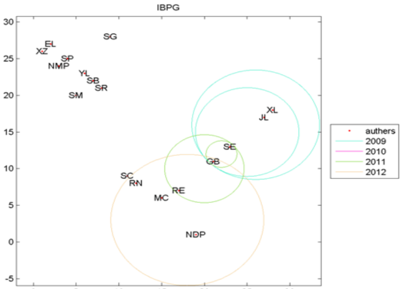


Figure 6. Papers that that share author G Bebek in the publication list

6. Conclusion

In this paper, we presented an approach that includes a number of spectral algorithms for visualizing overlapping sets using the eigenvectors of the Laplacian of two different graph representations: co-membership graph and bipartite graph. We first introduced an algorithm that is based on the co-membership graph of the sets to be visualized CMG. We have explored four different versions of the algorithm and showed that the weighted normalized version of the algorithm performs better than other versions. That is, because the normalization takes into account the degree of each node in the graph produced from the sets, and for this reason, treats each item equally (the degree of an item means that which items share membership with that particular item). Besides, when weights are considered, we actually are considering the number of times that any two items appear in the same set together, and this affects the relation between those two items. After introducing the CMG algorithm, another algorithm, i.e., the BPG algorithm has been presented. The power of BPG comes from the use of a bipartite graph instead of the co-membership graph, where the set membership information can be directly represented. Furthermore, using this model, the centers of the circles that we use to represent the sets can be directly computed.
Comprehensive experiments conducted and showed that the BPG outperforms CMG.

IBPG is a further improvement to BPG that optimizes radii of circles that represent sets, and it demonstrated better performance. However, the use of IBPG depends on whether or not the user wants to allow false negatives. Although IBPG produces some false negatives, it still provides what is required from a drawing algorithm: initial basic information about the structure of the dataset. Besides, the performance of the algorithm can be improved by feeding back subsets of data based on the outputs of the algorithm. Here, the input size will be smaller, allowing the algorithm to produce better looking results.

Finally, a discretization procedure is applied to IBPG to help isolating the items that lay over each other, so that they can be easily identified in the drawing.

A major limitation of the proposed method is the use of circles to represent the sets. A potentially useful improvement in this regard, would be to use the convex hull of the points that represent the items in a set. Since a convex shape would still cause errors, this can be further improved by refining the bounding shape by allowing non-convex shapes as well. However, such an approach would lead to more complicated optimization problems. Moreover, it is more beneficial to compute the center of the circle dynamically as we add items to the circle. The current approach computes the center first and then starts adding items and computes the candidate radii and errors. Since the centers are fixed, there is not much to do to keep the items close to the center. Computing the centers dynamically, though, allow us to cluster the items around the center and produce better drawing.

References

[1] M. Fisher and H. Gall. 2003. MDS-Views: Visualizing problem report data of large scale software using multidimensional scaling. In *Proceedings of ELISA*, Netherlands.

[2] J. Shi and J. Malik. 2000. Normalized Cuts and Image Segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intell.*, vol. 22 . pp. 888–905.

[3] B. Shneiderman, D. Feldman and A. Rose. 2000. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the ACM Digital Libraries Conference*. New York, ACM Press, pp. 57–66.

[4] J. Flower and J. Howse. 2002. Generating Euler Diagrams, Proc. Diagram, In *Proceedings of International Conference on the Theory and Application of Diagrams*. pp. 61-75.

[5] J. Flower, P. Rodger and P. Mutton. 2003. Layout Metrics for Euler Diagrams. In *Proceedings of 7th International Conference on Information Visualisation*. pp. 272-280.

[6] S. Chow and F. Ruskey. Drawing Area-Proportional Venn and Euler Diagrams. To appear in proceedings of GD2003. LNCS. Springer Velag.

[7] P. Simonetto, D. Auber, and D. Archambault. 2009. Fully automatic visualisation of overlapping sets. In *Computer Graphics Forum*. vol. 28(3), pp. 967–974.

[8] G. Di Battista, P. Eades, R. Tamassia and I.G. Tollis, Graph Drawing: Algorithms for the Visualization of Graphs, Prentice-Hall, 1999.

[9] M. Kaufmann and D. Wagner (Eds.), Drawing Graphs: Methods and Models, LNCS 2025, Springer Verlag, 2001.

[10] M. Frohlich and M. Werner. 1997. Demonstration of the interactive graph visualization system. da Vinci’, R. Tamassia and I. Tollis (eds.), Proc. In *Symp. Graph Drawing, GD ’94*, Lecture Notes in Computer Science, Berlin, Springer-Verlag. vol. 894, pp. 266–269.

[11] S. Clemencon, H. Arazoza, F. Rossi, and V. Tran. 2011. Hierarchical clustering for graph visualization. In *ESANN 2011*. Bruges, Belgium. pp. 227-232.

[12] Luo S, Liu C, Chen B, Ma K. Ambiguity-free edge-bundling for interactive graph visualization. IEEE Transactions on Visualization and Computer Graphics 2011; 99(PrePrints).

[13] K. Hall. 1970. An r-dimensional Quadratic Placement Algorithm. In *Management Science*, vol. 17, pp. 219-229.

[14] Y. Koren, L. Carmel and D. Harel. 2002. ACE: A Fast Multiscale Eigenvectors Computation for Drawing Huge Graphs. In *Proceedings of InfoVis’02*, IEEE, pp. 137–144.

[15] Y. Koren. 2003. On spectral graph drawing. In Computing and Combinatorics, In *Proceedings of COCOON 2003*, pp. 496–508.

[16] S. Roweis and L. Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding, *Science*, vol. 290.

[17] Tenenbaum, J. B., Silva, V. d., and Langford, J. C., 2000, A global geometric framework for nonlinear dimensionality reduction, Science, 290, 2319- 2323.

[18] Scholkopf, B., Smola, A. J., and M¨uller, K.-R., 1998, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation, 10, 1299-1319.

[19] J. de Leeuw and P. Mair. Multidimensional scaling using majorization: SMACOF in R. Technical Report 537, UCLA Statistics Preprints Series, 2009.

[20] A. Agarwal, J. Phillips and S. Venkatasubramanian. 2010. Universal multi-dimensional scaling. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '10). *ACM.* New York, NY, USA, pp. 1149-1158. DOI=10.1145/1835804.1835948

[21] L. Chen and A. Buja. 2009. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. In *Journal of the American Statistical Association*, vol. 104, pp. 209–219.

[22] Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0-a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24: 1650–1651.

[23] A. Abdalla. 2012. *SAFA: Spectral Approach for Automatic Set Visualization.* Master Thesis. Case Western Reserve University, USA.

[24] B. Alsallakh, W. Aigner, S. Miksch and H. Hauser. 2013. Radial Sets: Interactive Visual Analysis of Large

Overlapping Sets. In *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19(12), pp. 2496- 2505.

[25] W. Meulemans, N. Riche, B. Speckmann, B. Alper, T. Dwyer. 2013. KelpFusion: A Hybrid Set Visualization Technique. In *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19(11), pp. 1846-1858.

[26] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot and H. Pfister. 2014. UpSet: Visualization of Intersecting Sets. In *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20(12), pp. 1983-1992.

[27] M. Koyuturk. 2015. Department of Electrical Engineering & Computer Science، Case Western Reserve University. [Online]. Available: http://compbio.case.edu/koyuturk/publications/

[28] A. S. Abdalla, M. Koyuturk, A. M. Maatuk and A. O. Mohammed. Sets Visualization using their Graph Representation. In *the Proceedings of* ICEMIS '15, 6 pages. September 2015. DOI: http://dx.doi.org/10.1145/2832987.2833023.

Shortest Path of Metrics TYPE**g**=f(x,y)*(**dx**²+**dy**²)

Ahmad Tayyar[1]

Abstract— A geodesic is the real world analog of a straight line. Where a straight line on a flat piece of paper minimizes the distance between two points, a geodesic minimizes the distance between two points on any surface; be it flat or not.Supposing that we have a surface in spacegiven by the equation z = f (x, y).The search for a geodesic line on this surface, or more generally in the plan provided by an arbitrary metric, may be made by solving the coupled differential second orderequations of Euler-Lagrange system. More precisely, the search of the shortest path connecting two given points may be made by solving that system for a specific initial velocity.

In this paper we determine the geodesic lines corresponding the metric of type g =(dx2 + dy2) for f (x, y) defines positive.

Starting from a metric of this type, we determine the Euler-Lagrange system correspondence; its solutions are geodesics. We designed geodesics and the shortest path for the given metric and a specific function f (x, y).

We will need to determine the appropriate initial velocity for the system's numerical resolution of two differential equations of second order. Therefore, we are providing a suitable method for this.

Keywords— Euler-Lagrange, geodesics, metric, shortest path.

I. Introduction

Geodesic line of a surface is the entire trajectory of a moving particle connected to this surface without friction, and not subjected to any external force.

Suppose that the surface is expressed by the equation z = f (x, y). Then, the general metric form of this surface is:

$$g = Edx^2 + 2Fdxdy + Gdy^2`$$

Where the coefficients E, F, G are given by the formulas:

$$E = 1 + f_x^2 F = f_x f_y G = 1 + f_y^2$$

Where

$$f_x = \frac{\partial f}{\partial x}, f_y = \frac{\partial f}{\partial y}$$

Consider a metric of the form

$$g = \left(dx^2 + dy^2\right) \qquad (1-1)$$

Then, E= G= f(x,y), and F= 0.

The equations which determines geodesic lines can be obtained from the Lagrangian formulation; the Lagrangian of the metric (1-1) is

$$L = f(x,y)(\dot{x}^2 + \dot{y}^2)$$

Euler-Lagrange equations give

$$\begin{cases} \dfrac{d}{dt}\left(\dfrac{\partial L}{\partial \dot{x}}\right) - \dfrac{\partial L}{\partial x} = (2\dot{x}f(x,y))^{\cdot} - (\dot{x}^2 + \dot{y}^2)\dfrac{\partial f}{\partial x} = 0 \\ \dfrac{d}{dt}\left(\dfrac{\partial L}{\partial \dot{y}}\right) - \dfrac{\partial L}{\partial y} = (2\dot{y}f(x,y))^{\cdot} - (\dot{x}^2 + \dot{y}^2)\dfrac{\partial f}{\partial y} = 0 \end{cases}$$

$$2\ddot{x}f(x,y) + 2\dot{x}\left(\frac{\partial f}{\partial x}\dot{x} + \frac{\partial f}{\partial y}\dot{y}\right) - \dot{x}^2\frac{\partial f}{\partial x} - \dot{y}^2\frac{\partial f}{\partial y} = 0$$

$$2\ddot{x}f(x,y) + \dot{x}^2\frac{\partial f}{\partial x} + 2\dot{x}\dot{y}\frac{\partial f}{\partial y} - \dot{y}^2\frac{\partial f}{\partial x} = 0$$

$$2\ddot{x}f(x,y) + (\dot{x}^2 - \dot{y}^2)\frac{\partial f}{\partial x} + 2\dot{x}\dot{y}\frac{\partial f}{\partial y} = 0$$

$$2\ddot{y}f(x,y) + 2\dot{y}\left(\frac{\partial f}{\partial x}\dot{x} + \frac{\partial f}{\partial y}\dot{y}\right) - \dot{x}^2\frac{\partial f}{\partial y} - \dot{y}^2\frac{\partial f}{\partial y} = 0$$

$$2\ddot{y}f(x,y) + \dot{y}^2\frac{\partial f}{\partial y} + 2\dot{x}\dot{y}\frac{\partial f}{\partial x} - \dot{x}^2\frac{\partial f}{\partial y} = 0$$

$$2\ddot{y}f(x,y) + (\dot{y}^2 - \dot{x}^2)\frac{\partial f}{\partial y} + 2\dot{x}\dot{y}\frac{\partial f}{\partial x} = 0$$

The Euler-Lagrange equations corresponding to (1-2) are

$$\begin{cases} \ddot{x} = \dfrac{1}{2f(x,y)}\left[(\dot{y}^2 - \dot{x}^2)\dfrac{\partial f}{\partial x} - 2\dot{x}\dot{y}\dfrac{\partial f}{\partial y}\right] \\ \ddot{y} = \dfrac{1}{2f(x,y)}\left[(\dot{x}^2 - \dot{y}^2)\dfrac{\partial f}{\partial y} - 2\dot{x}\dot{y}\dfrac{\partial f}{\partial x}\right] \end{cases} \qquad (1-3)$$

II. Numerical solution for coupled differential second order equations

In order to solve the coupled differential equation (1-3), numerical method is used.The results are geodesics starting at an initial velocity from a specific point.To draw several geodesics starting from the same point, it would be enough to change the initial velocity.

The Runge-Kutta Method is effective for numerical integration of the coupled differential equations.

Supposing that we have coupled differential equations of the following form, we want to determine the points forming that solution.

[1] Ahmad Tayyar is an associate professor in Jerash University, Jordan (corresponding author’s phone: 00962795290252; e-mail: ahmad.tayyar@hotmail.com).

$$\begin{cases} \ddot{x} = f(x, y, \dot{x}, \dot{y}) \\ \ddot{y} = g(x, y, \dot{x}, \dot{y}) \end{cases}$$

To solve this statement, we convertit to first-order differential equations. For that, we set $\dot{x} = v \; and \; \dot{y} = w$ (New dependent variables).The coupled differential equations become:

$$\begin{cases} \dot{v} = f(x, y, v, w) \\ \dot{w} = g(x, y, v, w) \end{cases}$$

At a point Pn (xn, yn) of the solution, the next point Pn + 1 (xn + 1, yn + 1) is obtained by calculating the following:

$$x_{n+1} = x_n + h\{v_n + \frac{1}{6}(k_{11} + k_{21} + k_{31})\}$$

$$y_{n+1} = y_n + h\{w_n + \frac{1}{6}(k_{12} + k_{22} + k_{32})\}$$

$$v_{n+1} = v_n + \frac{1}{6}(k_{11} + 2k_{21} + 2k_{31} + k_{41})$$

$$w_{n+1} = w_n + \frac{1}{6}(k_{12} + 2k_{22} + 2k_{32} + k_{42})$$

Knowing that

$$k11 = hf(x_n, y_n, v_n, w_n)$$

$$k12 = hg(x_n, y_n, v_n, w_n)$$

$$k21 = hf\left(x_n + \frac{1}{2}hv_n + \frac{1}{8}hk_{11}, y_n + \frac{1}{2}hw_n + \frac{1}{8}hk_{12}, v_n + \frac{1}{2}k_{11}, w_n + \frac{1}{2}k_{12}\right)$$

$$k22 = hg\left(x_n + \frac{1}{2}hv_n + \frac{1}{8}hk_{11}, y_n + \frac{1}{2}hw_n + \frac{1}{8}hk_{12}, v_n + \frac{1}{2}k_{11}, w_n + \frac{1}{2}k_{12}\right)$$

$$k31 = hf\left(x_n + \frac{1}{2}hv_n + \frac{1}{8}hk_{11}, y_n + \frac{1}{2}hw_n + \frac{1}{8}hk_{12}, v_n + \frac{1}{2}k_{21}, w_n + \frac{1}{2}k_{22}\right)$$

$$k32 = hg\left(x_n + \frac{1}{2}hv_n + \frac{1}{8}hk_{11}, y_n + \frac{1}{2}hw_n + \frac{1}{8}hk_{12}, v_n + \frac{1}{2}k_{21}, w_n + \frac{1}{2}k_{22}\right)$$

$$k41 = hf\left(x_n + hv_n + \frac{1}{2}hk_{31}, y_n + hw_n + \frac{1}{2}hk_{32}, v_n + k_{31}, w_n + k_{32}\right)$$

$$k42 = hg\left(x_n + hv_n + \frac{1}{2}hk_{31}, y_n + hw_n + \frac{1}{2}hk_{32}, v_n + k_{31}, w_n + k_{32}\right)$$

III. EXAMPLE

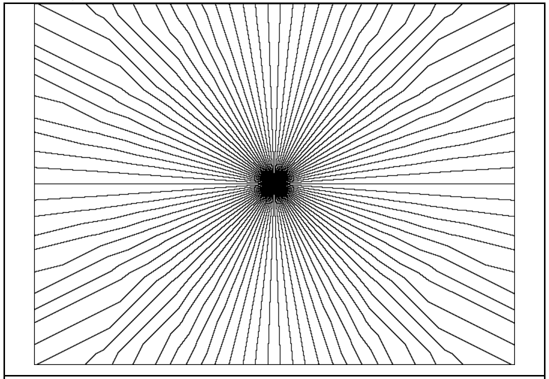
Consider

$$f(x, y) = x^2 + y^2 + 1$$

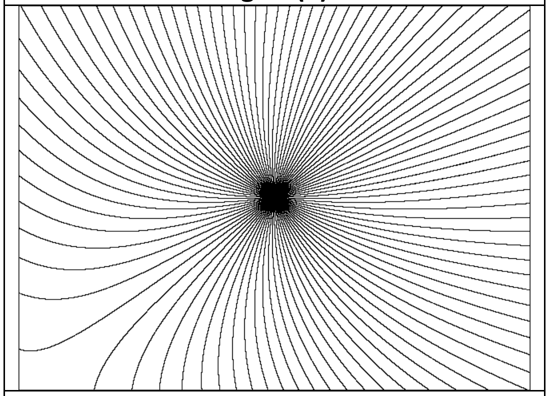
As per statement (1-3) becomes as follows:

$$\ddot{x} = \begin{cases} \frac{1}{2(x^2 + y^2 + 1)}[2x(\dot{y}^2 - \dot{x}^2) - 4y\dot{x}\dot{y}] \\ \ddot{y} = \frac{1}{2(x^2 + y^2 + 1)}[-4x\dot{x}\dot{y} + 2y(\dot{x}^2 - \dot{y}^2)] \end{cases} \qquad (3-1)$$

This is a system of second order which is dependent on both time and metric. Matlab can be used to solve this system using a numericalmethods. However, Runge Kutta method was implanted in C++ program in order to drawa geodesic for a given metric with initial conditions at different points of the plan XOY. This was applied by givingvalues for initial velocity ray angel [sin (θ), cos (θ)] for θ∈ [0, 2π], and with a small angular step of 4 degrees for example. Figures (1) and (2) illustrate the results for point O(0, 0) and point (x0, y0) = (320, 240), respectively.

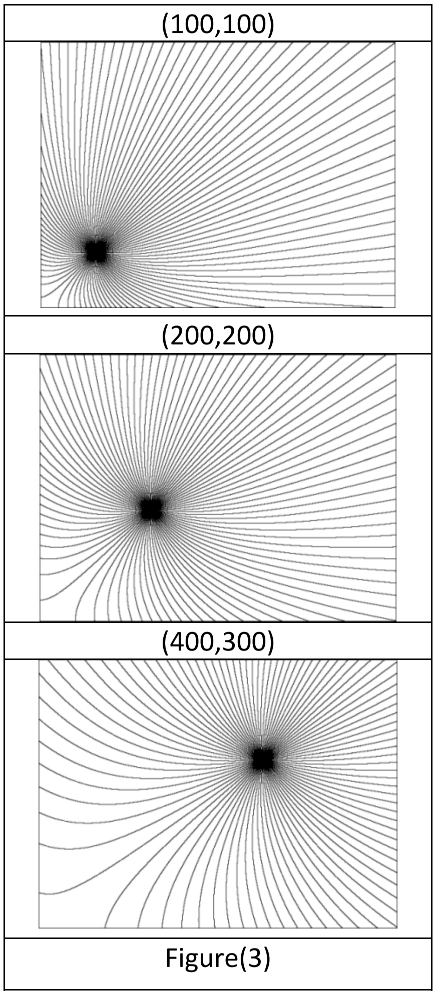


Figure(1)



Figure(2)

Furthermore, figure (3) shows the results at various other points.



(100,100)

(200,200)

(400,300)

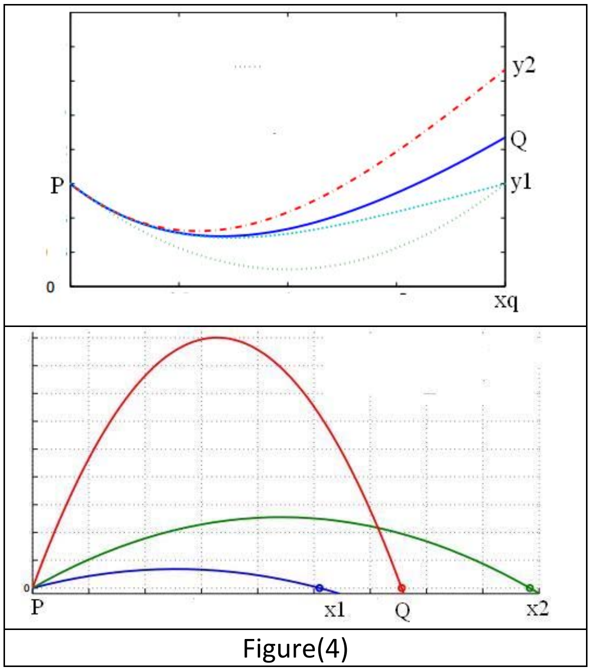
Figure(3)

IV. Searching for the proper initial velocity for coupled differential second order equations.

The main objective of this section is to provide a program to determine the initial velocity for a coupled differential equations when a numerical method is used to obtain a solution passes through two specific points.

Assuming that we have the coupled differential equations (1-3), we have to determine the proper initial velocity at one of two specific points to get a solution that pass the two points using Runge Kutta algorithm. For that purpose, Binary Search method is implemented.

Assuming that we need to find solution for the statement (3-1) passing through the points P(xn,yn) and Q(xq, yq).,we must find the value ofangle θ of the initial velocity ray applied at the point P: v0 (cosθ, sinθ) which passes from the point P to point Q. Suppose we spotted two solutions, one of them passes away from the point Q, and the second does not get to it. Here we distinguish two cases (Figure 4)



Figure(4)

A. *First Case*

The resulting solutions go through the same x-axis (xq) with different y-axis values. Suppose that two solutions are spottedin terms of y-axis sides for the point Q, as if we get a solution passes through the point (xq , y1) for angleθ1, and we get a solution passes through the point (xq, Y2) for the θ2 angle. In other words, y(θ1) = y1, y(θ2) = y2

Angle θ1 leads to the point P1(xq, y1), and angle θ2 leads to the point P2(xq, y2) and y1 ≤ yq ≤ y2. Then, we consider angle θ (θ = 0.5 (θ1 + θ2)) will give us a solution passing through a point with y-axis values as y(θ) = y0 and nearer to yq. Which means that:

$$dis(y_0 - y_q) < \min\left(dis(y_1,y_q), dis(y_2,y_q)\right)$$

So the angle θ gives a solution that converges more to the point Q(xq, yq).

We repeat using the angles (θ1, θ) if y(θ1)= y0<yq. However, we repeat using the angles (θ, θ2) if y(θ1)= y0 > yq.

We repeat this recursive functionuntil y(θ) becomes close enough to yq. This means getting a solution (almost) passes through the point Q(xq, yq).

This method is used to determine the initial velocity angle that converges to point Q(xq, yq) because y1 and y2 are located on different sides of Q(xq, yq)

Note:

This method of scanning all the possible directions for the initial velocity ray that is applied in the starting point, will not be efficient when we study coupled differential second orderequations that are so sensitive to the initial velocity. Then, we will not be able to find a proper initial velocity to find a solution that connects two fixed points.

B. Second Case

The resulting solutions go through the same y-axis (yq) with different x-axis values. Suppose that two solutions are spottedin terms of x-axis sides for the point Q, as if we get a solution passes through the point (x1 , yq) for angle θ1, and we get a solution passes through the point (x2, yq) for the θ2 angle. In other words, $x(\theta 1) = x1$, $x(\theta 2) = x2$.

Angle θ1leadsto the point Q1(x1, yq), and angle θ2 leads to the point Q2(x2, yq) and $x1 \leq xq \leq x2$.

Doing the same previous reasoning in (4.1); we repeat with the angles (θ1, θ) if $x(\theta 1)= x0 <xq$. However, we repeat with the angles (θ, θ2) if $x(\theta 1)=x0 > xq$.

Depending on that, we can present the following recursive function, knowing that RK2D(teta,x0) is the function that is related to Runge-Kutta algorithm.

```
void velocity0(float teta1,float teta2, float y1, float y2)
```

When applying the previous function with the statement (3-1), we get figure(5) on the computer’s screen for $\theta1=-\pi/4$, $\theta2=\pi/2$. That was for a geodesic line connecting the two points P1(100, 300) and P2(400, 300). We found that proper angle for the initial velocity is $\theta \approx 0.4786$.

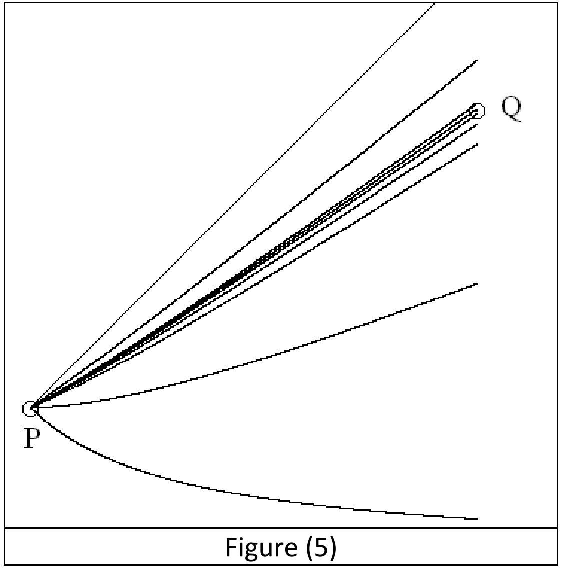


Figure (5)

V. Conclusion

We showed how to compute a geodesic corresponding to a metric, initial position and velocity.

Therefore, we can replace the function f(x, y) in the metric (1-1) with a function of our choice.Therefore, we get a specific measurement and coupled differential second orderequations of Euler-Lagrange system corresponding to that measurement. Usually, we will have to use a numerical method to solve it. The described method in paragraph 4 is useful to determine the proper initial velocity to choose the geodesic line that passes two specific points for the studied measurement.

References

[1] D. Lehmann et C. Sacré
Géométrie et topologie des surfaces
Edition BUF, 1982

[2] AKAI, TERRENCE J., Applied Numerical Methods for Engineers, Wiley, 1994, ISBN 0-471-57523-2.

[3] J.M. Beck, R.T. Farouki et J. k. Hinds
Surface analysis methodes
I.E.E.E., pp 18-36, 1986.

[4] M.P. Do Carmo
Differential geometry of curves surfaces
Prentice Hall, Englewood Cliffs, N. J., 1976

[5] R.S. Millman et G.D. Parker
Elements of differential geometry
Prentice-Hall.inc, Englewood Cliffs, New Jerzey.

[6] T.J. Willmore
An introduction to differential geometry
Oxford University Press, Oxford

An Innovative Imputation and Classification Approach for Accurate Disease Prediction

Yelipe UshaRani
Department of Information Technology
VNR VJIET
Hyderabad, INDIA

Dr.P.Sammulal
Dept.of Computer Science and Engineering
JNT University
Karimnagar, INDIA

***Abstract*—Imputation of missing attribute values in medical datasets for extracting hidden knowledge from medical datasets is an interesting research topic of interest which is very challenging. One cannot eliminate missing values in medical records. The reason may be because some tests may not been conducted as they are cost effective, values missed when conducting clinical trials, values may not have been recorded to name some of the reasons. Data mining researchers have been proposing various approaches to find and impute missing values to increase classification accuracies so that disease may be predicted accurately. In this paper, we propose a novel imputation approach for imputation of missing values and performing classification after fixing missing values. The approach is based on clustering concept and aims at dimensionality reduction of the records. The case study discussed shows that missing values can be fixed and imputed efficiently by achieving dimensionality reduction. The importance of proposed approach for classification is visible in the case study which assigns single class label in contrary to multi-label assignment if dimensionality reduction is not performed.**

Keywords— imputation; missing values; prediction; nearest neighbor, cluster, medical records, dimensionality reduction

I. Introduction

Medical records preprocessing is an important step which cannot be avoided in most of the situations and when handling medical datasets. The attributes present in medical records may be of different data types. Also, the values of attributes have certain domain which requires proper knowledge from medical domain to handle them.

This is because of this diverse nature of medical records, handling medical records is quite challenging for data miners and researchers. The various preprocessing techniques for medical records include fixing outliers in medical data, estimation and imputing missing values, normalizing medical attributes, handling inconsistent medical data, applying smoothing techniques to attributes values of medical records to specify some of them.

Data Quality depends on Data Preprocessing techniques. An efficient preprocessing of medical records may increase the data quality of medical records. In this context, data preprocessing techniques have achieved significant importance from medical data analysts and data miners. Incorrect and improper data values may mislead the prediction and classification results, there by resulting in false classification results and thus leading to improper medical treatment which is a very dangerous potential hazard. This research mainly aims at handling missing attribute values present in medical records of a dataset. The attributes may be numeric, categorical etc. The present method can handle all the attribute types without the need to devise a different method to handle different attribute types. This is first importance of our approach. We outline research objective and problem specification in the succeeding lines of this paper and then discuss importance of our approach.

A. Research Objective

We have the following research objectives in this research towards finding missing values

- Obviously our first and foremost objective is to impute missing values.
- Aim at dimensionality reduction process of medical records.
- Classify new medical records using the same approach used to find missing values.
- Cluster medical records to place similar records in to one group.

B. Problem Specification

Given a dataset of medical records with and without missing values, the research objective is to fix set of all missing values in the medical records by using a novel efficient Imputation approach based on clustering normal medical records, so as to estimate missing values in medical records with missing values.

C. Importance of Present Approach

The importance of the present approach which we wish to propose has the following advantages

- The method may be used to find missing attribute values from medical records
- The same approach for finding missing values may be used to classify medical records
- The disease prediction may be achieved using the proposed approach without the need to adopt a separate procedure

- Handles all attribute types
- Preserves attribute information
- May be applied for datasets with and without class labels which is uniqueness of the current approach.

II. Related works

Most of the research works carried in the literature argues that the presence of missing values of medical attributes makes the extra overhead may be in prediction and classification or when performing dataset analysis. In contrast to this the researchers Zhang, S et.al in their work [1] discuss and argue that missing values are useful in cost sensitive environments [17-18]. This is because some of the attributes values incur high cost to fill those values by carrying experiments. In such cases, it would be cost effective to skip such tests and values associated with those medical tests. Handling Missing values in medical datasets is quite challenging and also requires use of statistical approaches [15, 16] to estimate the same. In [2], missing values are found by using clustering approach where the missing value is filled with the value of attribute of nearest clusters. The concept of support vector regression and clustering is applied to find missing values in the work of authors [3]. In [4], phylogeny problems occurring because of missing values is discussed. An approach to handle medical datasets consisting mixed attribute types is handled in [5]. Some of the research contributions in missing values include [6-21].

III. Research Issues in Mining Medical Data

A. Handling Medical Datasets

The research should first start with the studying the benchmark datasets. Sometimes there may be a need to start collecting data from scratch if we are working over a problem in particular domain. However, when working with medical datasets, we need to remember that the dataset is multi-variate.

B. Handling Missing Values in Medical Datasets

The medical datasets are not free from missing values. Obviously there is no free lunch. We must make sure to handle the missing values suitably and accurately. A simple approach would be to discard the whole record which essentially contains the missing value of an attribute. Some significant novel approaches include [6-8, 12-14].

C. Choice of Prediction and Classifications algorithms

The underlying dataset is the deciding factor for choice of the algorithm. A single classification algorithm is not suitable for every dataset. Recent works include [18-19].

D. Finding Nearest Medical Record and Identifying the Class Label

The heart of any classification or clustering algorithm is the distance measure used to estimate the record distance between any two records. Since classification involves training and testing phases, training dataset must include all possible combinations which forms a knowledge database using which class label is estimated accurately. Finding nearest records may be performed through using KNN-classifiers or using any other classifiers. Classification may has a curse of dimensionality. Hence, dimensionality reduction must be suitably addressed. However, this can make situation complex and also inaccurate sometimes, if important attribute information or attributes are missed or discarded [9-11].

E. Deciding on Medical Attributes

The attributes of the medical dataset are also the prime concern in prediction and classification. This is because the attributes are multi variants [11]. Coming up with the deciding attributes for heart disease prediction which can make significant impact on the classification accuracy and prediction of the disease symptom is also one of the important tasks. In short, it is required to perform a thorough literature survey, fix the attributes which must be considered and which may be discarded.

F. Removal of Noise

After deciding the number of attributes, we may have one or more attributes which may be not important and hence may be discarded without any loss. Every effort must be made in this direction, so that the attributes which are of least importance and removal of the attributes does not make any significant affect may be eliminated.

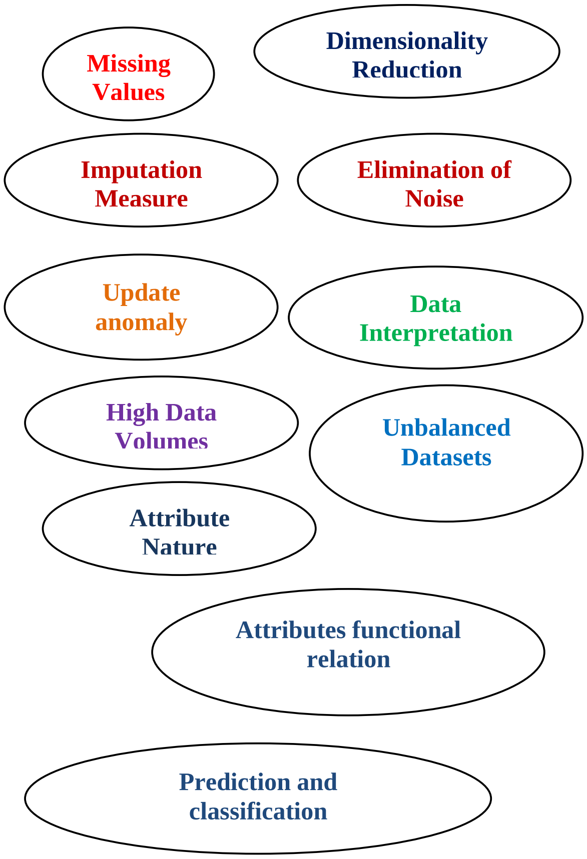


Fig. 1 Research Problems when handling Medical Datasets

IV. IMPUTATION FRAME WORK

In this section, we discuss framework to impute missing values as shown in Fig.2 and Fig.3

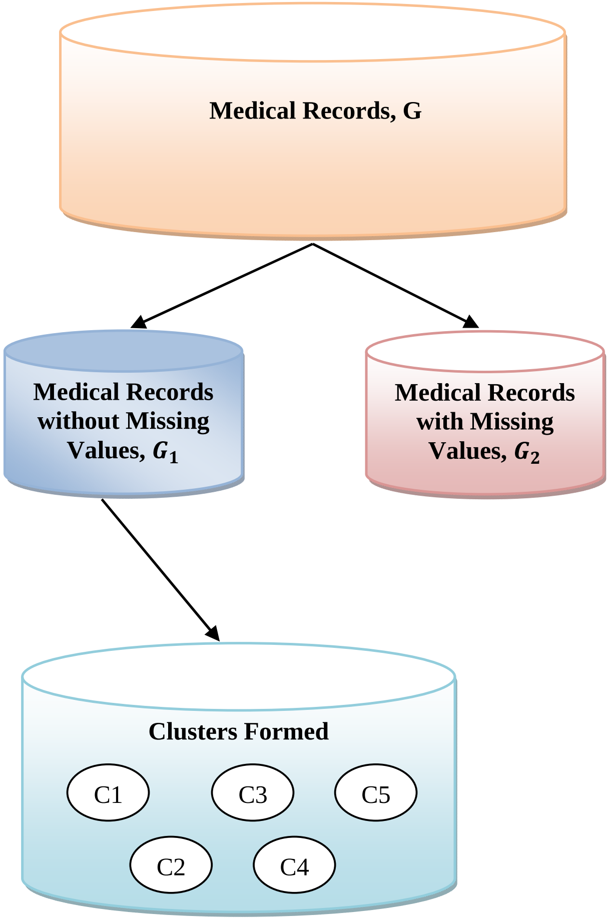

Medical Records, G
Medical Records without Missing Values, G_1
Medical Records with Missing Values, G_2
Clusters Formed
C1
C3
C5
C2
C4

Fig. 2 Generating Clusters from medical records

The framework for missing value Imputation consists of following steps. The approach for missing values is based on the concept of clustering medical records without missing values. This is because, all similar medical records shall come into one cluster and hence imputation performed shall be more accurate. This approach of finding missing values has not been carried out earlier in the literature. We present analytical framework with a case study in this paper. This research was motivated form the work by researchers for intrusion detection published in 2015 [20].

A. Generating Clusters from Group G_1

- This step involves finding the number of class labels and generating number of clusters equal to number of class labels
- The clusters may be generated using k-means algorithm by specifying value of k to be number of class labels. Alternately, we may apply any clustering algorithm which can generate k clusters

B. Computing distance of normal records to Cluster Centers

- Obtain mean of each cluster. This shall be the cluster center
- Obtain distance of each medical record to each cluster center.
- Sum all distances obtained
- The result is all medical records mapped to single value achieving dimensionality reduction.

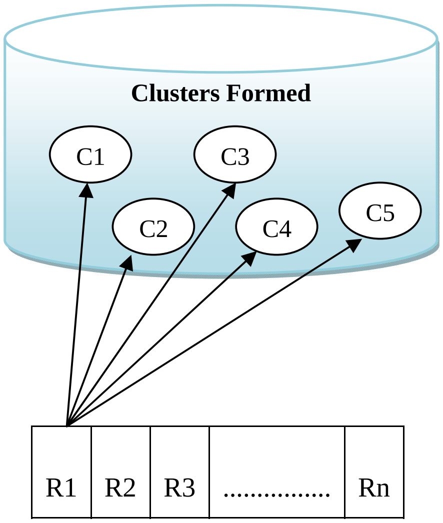


Fig. 3 Distance Computation from Record to Cluster Centers

C. Computing distance of missing records to Cluster Centers

- Obtain distance of each medical record having missing values to each cluster center by discarding those attributes with missing values.
- Sum all distances obtained
- The result is all medical records mapped to single value achieving dimensionality reduction.
- Method preserves information of attributes

D. Find Nearest Record to Impute Missing Values

Consider each missing record in group, G_2 one by one. Find the distance of this record to all the records in group G_1.The record to which the distance is minimal, shall be the nearest neighbor. Perform imputation of the missing attribute value by considering the corresponding attribute value of nearest record in that class. The frequency may also be considered for imputation incase, we have more than one nearest neighbors.

V. Proposed Imputation Algorithm

A. Proposed Algorithm

$\boldsymbol{Input}$: $Medical\ Records\ with\ Missing\ Values$

$\boldsymbol{Output}$: $Imputation\ of\ Missing\ Values$

$\boldsymbol{Notations\ adopted}$:

R_i $-$ i^{th} $medical\ record$

$R_i(A_K)$ $-$ k^{th} $attribute\ value\ of\ i^{th} medical\ record$

G_c $-$ c^{th} $group$

i, k $-$ $index\ of\ medical\ records\ and\ attributes$

$\emptyset$ $-$ $misisng\ record\ or\ Empty\ record\ value$

c $-$ $number\ of\ decision\ classes\ in\ medical\ dataset$

D_d $-$ $d^{th} decision\ class$

m $-$ $total\ number\ of\ medical\ records$

n $-$ $number\ of\ attributes\ in\ each\ record$

μ_d $-$ $cluster\ center\ of\ d^{th}\ cluster$

μ_{dn} $-$ $mean\ value\ of\ n^{th}\ attribute$

h $-$ $number\ of\ records\ in\ group\ , G_2$

$z\ -\ number\ of\ records\ in\ group\ , G_1\ equal\ to\ (m-h)$

Step-1: Read Medical Dataset

Read the medical dataset consisting of medical records. Find records with and without missing values. Classify records in to two groups, say G_1 and G_2. The first group, G_1 is set of all medical records with no missing values given by Eq.1. The second group, G_2 is set of all medical records having missing values given by Eq.2.

$$G_1 = U\{R_i \mid R_i(A_K) \neq \emptyset, \forall\, i,k\} \quad (1)$$

$$G_2 = U\{R_i \mid R_i(A_K) = \emptyset\ / \exists\, i,k\} \quad (2)$$

Where $i \in (1, m-h)$ and $k \in (1,n)$.We may consider group, G_1 as training set of medical records while group, G_2 is considered as testing set in this case.

Step-2: Cluster Medical Records with No Missing values

Let, g = $|D_d|$, be the number of decision classes. Determine the maximum number of decision classes available in the medical dataset being considered. Cluster the medical records in group, G_1 to a number of clusters equal to g. i.e $|D_d|$.

This may be achieved using K-means clustering algorithm. This is because K-means algorithm requires the number of required clusters to be specified well ahead before clustering process is carried out. The output of step-2 is a set of clusters. i.e Number of output clusters is equal to‘g’.

This is shown in fig.4 and fig.5 where a set of medical records represented by $G_1$are clustered in to ‘d’ clusters.

Step-3: Obtain Cluster Center for each Cluster formed

This step involves finding the cluster center for each cluster which is generated using the k-means clustering algorithm. We can obtain the cluster center by finding the mean of each attribute from attribute set, A_K of medical attributes.

Let Cluster- C_d denotes d^{th} cluster having the records R_1, R_6, R_8 and R_9 with single attribute. Then the cluster center is given by Eq.3 as

$$\mu_d = \frac{R_1(A_1) + R_6(A_1) + R_8(A_1) + R_9(A_1)}{4} \quad (3)$$

In general the cluster center of g^{th} cluster may be obtained using the generalized equation, Eq.4 given below

$$\mu_g = U_k\left[\frac{\{\sum R_l^k \mid l \in \{1,q\} \text{ for each } k \in \{1,n\}\}}{|l|}\right] \quad (4)$$

μ_g is hence a sequence of ‘n’ values indicating cluster center over ‘n’ attributes. The notation, U_k is used to denote set of all values each separated by a symbol comma. The cluster center may hence be formally represented using the representation given by Eq.5

$$\mu_g = <\mu_{g1}, \mu_{g2}, \mu_{g3}, \mu_{g4}, \ldots \ldots . \mu_{gn} > \quad (5)$$

Here ‘n’ indicates total number of attributes in each medical record and |g| indicates number of clusters.

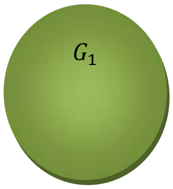


Fig 4. G_1 Before Clustering

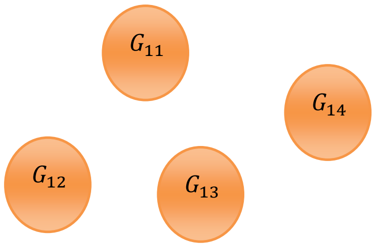


Fig. 5 Before and After Clustering

Step-4: Compute distance between each $\boldsymbol{R_i}$ and each $\boldsymbol{\mu_d}$

Find distance from each medical record, $\boldsymbol{R_i}$ in group, $G_1$to each of the cluster centers, $\boldsymbol{\mu_d}$ obtained in step-3. This can be achieved through finding the Euclidean distance between each medical record of ‘n’ attributes and cluster center of each cluster defined over 'n' attributes. These cluster distances computed are summed to obtain a single distance value. This distance is called Type-1 distance value given by Eq. 6.

$$\mathrm{Dist}^{\mathrm{d}}(\boldsymbol{R_i}, \boldsymbol{\mu_d}) = \sqrt{(R_{i1} - \mu_{d1})^2 + (R_{i2} - \mu_{d2})^2 + \cdots (R_{in} - \mu_{dn})^2} \quad (6)$$

$$\forall\, i \in (1, n), \forall d$$

At the end of Step-4 we have distance value from each record , R_i to each cluster center denoted by μ_d.

Step-5: Transform multi-dimensional medical record to a single dimension numeric value by using mapping function

$Map(R_i)$ is a mapping function which maps the medical record, R_i to a single distance value. To determine mapping function value of a record we use the equation, Eq.7

This can be obtained by adding all distances obtained in Step-4

$$Map(R_i) = \sum_{d=1}^{d=|g|} \sum \mathrm{Dist}^{\mathrm{d}}(R_i, \mu_d) \qquad \forall\, i \in (1, n-h) \quad (7)$$

Where |g| is number of clusters formed, (n-h) indicates number of records in group, G_1.

At the end of step-5, we have each medical record, R_i mapped to a single dimension distance value. In other words, the medical record of 'n' dimensions is reduced to single dimension achieving dimensionality reduction.

Step-6: Compute distance between each $\boldsymbol{R_j}$ in group, $\boldsymbol{G_2}$ and each $\boldsymbol{\mu_d}$ of clusters formed

Obtain distance value of missing records to these cluster centers by discarding the attributes with missing values. Find distance from each medical record, $\boldsymbol{R_j}$ in group, G_2 to each of the cluster centers, $\boldsymbol{\mu_d}$ obtained in step-3.

This can be achieved through finding the Euclidean distance between each medical record of 'n' attributes and cluster center of each cluster defined over 'n' attributes. These cluster distances computed are summed to obtain a single distance value. This distance is called Type-2 distance value given by Eq.8.

$$\mathrm{Dist}^{2}\,(\boldsymbol{R_{in}}, \boldsymbol{\mu_{dn}}) = \sqrt{\sum (R_{i1} - \mu_{d1})^2 \ldots\ldots\ldots.} \forall R_{in}\ where\ n \neq y \quad (8)$$

discarding the y^{th} missing attribute value.

Step-7: Transform multi-dimensional medical record with missing values to a single dimension by using mapping function

$Map^r(R_j)$ is a mapping function which maps each medical record, R_j consisting to a single distance value. To determine mapping function value of a record R_j , we use the equation, Eq.9

$$Map^{r}(R_j) = \sum_{d=1}^{d=|g|} \sum \mathrm{Dist}^{\mathrm{d}}(R_j, \mu_d) \qquad \forall\, j \in (1, h) \quad (9)$$

Where |g| is number of clusters formed and $j \epsilon (1, h)$

At the end of step-7, we have each medical record, R_j mapped to a single dimension distance value. In other words, the medical record of 'n' dimensions is reduced to single dimension achieving dimensionality reduction.

Step-8: Obtain difference between distances obtained in step-6 and step-8

For each missing record, R_j in G_2, obtain difference between mapping functions of each record, R_i in group, G_1 and missing record, R_j in group, G_2. Call this value as d_{ij}

Step-9: Find nearest record

The medical record R_j is most similar to the medical record, R_i whose corresponding d_{ij} is most minimum as given by Eq.10.

$$d_{ij} = Min_i \{Map(R_i) - Map^{r}(R_j)\} \qquad (10)$$

Step-10: Fix Missing values and Impute Missing Values

The medical record R_j is most similar to the record, R_i whose corresponding d_{ij} is most minimum. In this case, impute the missing attribute value of record, R_j denoted by R_{jr} by the attribute value, R_{ir} of medical record denoted as R_i.

Incase more than one record with same minimum value is obtained then, fill the missing value of the attribute with the attribute value whose frequency is maximum from the same decision class. Alternately, we may fix the mean of the values also from the corresponding decision class attribute values.

VI. Case Study

In this Section-VI, we discuss case study to find missing attributes values of medical records by using the proposed approach. For this, we consider a sample dataset consisting sample values.

Consider Table. I, shown below consisting of sample dataset of medical records having categorical and numerical values. Table. II shows medical records without missing values after normalizing sample dataset. Table.III denotes records with and without missing values. Table IV denotes all records without missing values and Table. V shows records with missing attribute values.

Table.VI depicts clusters generated from group G_1 , which consists medical records with no missing values after applying k-means algorithm. There are two clusters generated C_1 and C_2.

C_1 contains set of all medical records {R_1,R_4,R_6,R_9} and C_2 contains set of all medical records {R_2, R_7, R_8 }. Table.VII gives the distances of records in group, G_1 to cluster center of the first cluster. Similarly, Table.VIII gives the distances of records in group, G_2 to cluster center of the second cluster.

Table.IX depicts computation values of mapping function of records of group, G_1. The mapping function $Map(R_i)$ is mapping distance of i^{th} record, which is sum of all distances from record, R_i to each of those cluster centers generated from application of clustering algorithm.

Table. X gives the distances of medical records in group, G_2 to each of the cluster centers.

Table. XI depicts computation values of mapping function of medical records containing missing values of group, G_2. The mapping function $Map^r(R_j)$ is mapping distance of jth record, which is sum of all distances from record, R_j to each of those cluster centers generated from application of clustering algorithm. The distance is computed considering those attributes which do not have missing values. i.e Attribute values are defined and recorded.

TABLE I. DATASET OF MEDICAL RECORDS

Record	A1	A2	A3	A4	Decision Class
R1	c_{11}	5	d_{31}	10	CLASS-1
R2	c_{13}	7	d_{31}	5	CLASS-1
R3	c_{11}	7	d_{32}	7	CLASS-1
R4	c_{12}	5	d_{31}	10	CLASS-1
R5	c_{13}	3	d_{32}	7	CLASS-2
R6	c_{12}	9	d_{31}	10	CLASS-2
R7	c_{11}	5	d_{32}	3	CLASS-2
R8	c_{13}	6	d_{32}	7	CLASS-2
R9	c_{12}	6	d_{32}	10	CLASS-2

TABLE II. NORMALIZED DATASET OF RECORDS

Record	A1	A2	A3	A4	Decision Class
R1	1	5	1	10	CLASS-1
R2	3	7	1	5	CLASS-1
R3	1	7	2	7	CLASS-1
R4	2	5	1	10	CLASS-1
R5	3	3	2	7	CLASS-2
R6	2	9	1	10	CLASS-2
R7	1	5	2	3	CLASS-2
R8	3	6	2	7	CLASS-2
R9	2	6	2	10	CLASS-2

TABLE III. RECORDS WITH AND WITHOUT MISSING VALUES

Record	A1	A2	A3	A4	Decision Class
R1	1	5	1	10	CLASS-1
R2	3	7	1	5	CLASS-1
R3	1	7	NaN	7	CLASS-1
R4	2	5	1	10	CLASS-1
R5	3	3	2	NaN	CLASS-2
R6	2	9	1	10	CLASS-2
R7	1	5	2	3	CLASS-2
R8	3	6	2	7	CLASS-2
R9	2	6	2	10	CLASS-2

TABLE IV. RECORDS WITH OUT MISSING VALUES

Record	A1	A2	A3	A4	Decision Class
R1	1	5	1	10	CLASS-1
R2	3	7	1	5	CLASS-1
R4	2	5	1	10	CLASS-1
R6	2	9	1	10	CLASS-2
R7	1	5	2	3	CLASS-2
R8	3	6	2	7	CLASS-2
R9	2	6	2	10	CLASS-2

TABLE V. RECORDS WITH MISSING VALUES

Record	A1	A2	A3	A4	A5	Decision Class
R3	1	7	?	7	1	CLASS-1
R5	3	3	2	?	3	CLASS-2

TABLE VI. CLUSTERS GENERATED USING K-MEANS

Clusters	Medical Records with out missing values
C1	R1,R4,R6,R9
C2	R2,R7,R8

TABLE VII. DISTANCE OF RECORDS TO CLUSTER-1

Record	Distance to Cluster-1
R1	5.312459
R2	1.374369
R4	5.153208
R6	5.878397
R7	2.624669
R8	2.134375
R9	5.022173

TABLE VIII. DISTANCE OF RECORDS TO CLUSTER-2

Record	Distance to Cluster-2
R1	1.47902
R2	5.214163
R4	1.299038
R6	2.772634
R7	7.189402
R8	3.344772
R9	0.829156

TABLE IX. MAPPING DISTANCE OF MEDICAL RECORDS WITHOUT MISSING VALUES

Record	Mapping Distance
R1	6.791479
R2	6.588532
R4	6.452246
R6	8.651031
R7	9.814071
R8	5.479147
R9	5.851329

TABLE X. MISSING VALUE RECORD DISTANCES TO CLUSTERS

Record	Distance to Cluster-1	Distance to Cluster-2
R3	2.603417	3.181981
R5	3.091206	3.561952

TABLE XI. MAPPING DISTANCE OF MEDICAL RECORDS WITH MISSING VALUES

Record	Mapping Distance
R3	6.791479
R5	6.588532

TABLE XII. DISTANCE OF MEDICAL RECORDS R3 WITH OTHER RECORDS

Record	Distance with R3
R1	0
R2	-0.20295
R4	-0.33923
R6	1.859552
R7	3.022592
R8	-1.31233
R9	-0.94015

Table.XII shows distance of record, R_3 to records R_1, R_2, R_4, R_6, R_7, R_8, R_9 . The record R_3 is nearest to medical record R_8. The Table. XIII shows nearest medical record R_8 for R_3 which is ideal record to carry imputation. The attribute value to be imputed is 2. i.e the categorical attribute value $\mathbf{d_{32}}$**.** This is because the attribute value, $\mathbf{d_{32}}$ was mapped to numerical value **2.**

TABLE XIII. NEASREST MEDICAL RECORD FOR RECORD R8

Record	A1	A2	A3	A4	Decision Class
R8	3	6	2	7	CLASS-2

TABLE XIV. DISTANCE OF MEDICAL RECORD R5 TO OTHER RECORDS

Record	Distance with R5
R1	0.202947
R2	0
R4	-0.13629
R6	2.0625
R7	3.225539
R8	-1.10939
R9	-0.7372

TABLE XV. NEASREST MEDICAL RECORD FOR RECORD R5

Record	A1	A2	A3	A4	Decision Class
R8	3	6	2	7	CLASS-2

Table.XIV shows distance of record, R_5 to records R_1, R_2, R_4, R_6, R_7, R_8, R_9 . The record R_5 is nearest to medical record R_8. The Table. XV shows nearest medical record R_8 for R_5 which is ideal record to carry imputation. The attribute value to be imputed is 7. i.e the numerical value**.**

Finally in this case study, we fill the missing values of medical records by imputing the missing attribute values. Since, attribute values after imputing, happen to be the same values which were present initially in Table.1 the correctness of the approach can be verified and validated. The proposed approach of finding imputation values is hence accurate and also efficient as it also aims at dimensionality reduction of medical records and then estimates missing values. In the process of dimensionality reduction we never miss any attribute values or attributes. This brings the accuracy in the present approach.

This approach may be extended to classify new medical record without class label to an appropriate class, if required by simply assigning class label of medical record to which the new record distance minimum. In this way, disease prediction or classification may be achieved.

VII. CLASSIFICATION OF NEW MEDICAL RECORDS

Consider the table of medical records with class labels as in Table.XVI with the parameter values same as Table. II, the last column is decision class, which predicts the disease level or stage. This table is free from missing values and hence is suitable for mining medical records.

TABLE XVI. NORMALIZED MEDICAL RECORDS WITH CLASSES

Record	P1	P2	P3	P4	Disease Class or Type
R1	1	5	1	10	Level-1
R2	3	7	1	5	Level-1
R3	1	7	2	7	Level-1
R4	2	5	1	10	Level-1
R5	3	3	2	7	Level-2
R6	2	9	1	10	Level-2
R7	1	5	2	3	Level-2
R8	3	6	2	7	Level-2
R9	2	6	2	10	Level-2

Assume that, we have an incoming medical record with the attribute values as R10 = [2, 5, 2, 9]. We can obtain Euclidean distances from record R10 to all the records R1 through R9. The class of the record is the class of medical record to which the Euclidean distance is minimum. Table. XVII gives distance of medical record R10 to all records.

TABLE XVII. NORMALIZED MEDICAL RECORDS WITH CLASSES

Record	Distance with R10
R1	1.732051
R2	4.690416
R3	3
R4	1.414214
R5	3
R6	4.242641
R7	6.082763
R8	2.44949
R9	1.414214

Using this approach we get two class labels as the record is nearest to both records R4 and R8. But the classes are class-1 and classs-2 for R4 and R8 respectively. So we can't categorize the disease correctly or accurately. This is because; we did not perform dimensionality reduction. This is overcome if we extend the approach for fixing missing values to classification also. The only difference is that we continue to extend the procedure outlined in Section –IV, to all the records after fixing missing values (R1 to R9) and adopt the procedure for missing record to the new record but considering all attribute values. This is shown in computations below.

Table. XVIII shows clusters generated using k-means with K=2. Table.XIX and Table.XX shows distance of records R1 to R9 w.r.t cluster centers. Table XXI and Table XXIII shows mapping distance of R1 to R9 and Record R10 respectively.

TABLE XVIII. CLUSTERS GENERATED USING K-MEANS

Clusters	Medical Records with out missing values
C1	R1,R4,R6,R9
C2	R2,R7,R8,R3,R5

Table.XXII gives distance value of new records R10 to clusters formed. Table. XXIV gives difference of mapping values of existing records and new record.

TABLE XIX. DISTANCE OF RECORDS TO CLUSTER-1

Record	Distance to Cluster-1
R1	1.47902
R2	5.214163
R3	3.269174
R4	1.299038
R5	4.656984
R6	2.772634
R7	7.189402
R8	3.344772
R9	1.47902

TABLE XX. DISTANCE OF RECORDS TO CLUSTER-2

Record	Distance to Cluster-2
R1	4.481071
R2	1.969772
R3	2.209072
R4	4.322037
R5	2.979933
R6	5.46626
R7	3.11127
R8	1.509967
R9	4.481071

TABLE XXI. MAPPING DISTANCE OF MEDICAL RECORDS

Record	Mapping Distance
R1	5.960091
R2	7.183935
R3	5.478246
R4	5.621075
R5	7.636917
R6	8.238894
R7	10.30067
R8	4.854739
R9	5.960091

TABLE XXII. DISTANCES OF NEW MEDICAL RECORD TO CLUSTERS

Record	Distance to Cluster-1	Distance to Cluster-2
R10	1.785357	3.628027

TABLE XXIII. MAPPING DISTANCE OF R10

Record	Mapping Distance
R10	5.053384

TABLE XXIV. DIFFERENCE OF MAPPING DISTANCE OF NEW MEDICAL RECORD TO MAPPING DISTANCE OF EXISTING RECORDS

Record	Distance with R3
R1	0.906707
R2	2.130551
R3	0.424862
R4	0.567691
R5	2.583533
R6	3.18551
R7	5.247288
R8	-0.19865
R9	0.906707

The class label of record R10 to be assigned is the class label of the record to which the distance is minimum in Table.XXIV. In this case, the distance of R10 is proved to be minimum w.r.t R8 as compared to other record distances. Hence the class label of the new record R10 is class label of record R8. i.e Class-2 or Level-2.

Hence, the category of disease level of person whose medical record values are defined by R10 is level-2 or class-2 or Type-2 disease.

VIII. ADVANTAGE OF PROPOSED METHOD

If we can see the result obtained with traditional approach without dimensionality reduction carried out, the category of disease is either Class-1 or Class-2. This is because of noise attribute values. We overcome such a disadvantage using the proposed method of dimensionality reduction and classification. Using this proposed method, we get a single class label for the new record. In our case, the class is identified as Class-2 or Level-2 using proposed method of classification. This is because of dimensionality reduction performed to single value without missing any attribute or neglecting any attribute value.

IX. DISCUSSIONS AND OUTCOMES

In this research, we address the first challenge of handling medical records in datasets. We discuss the approach for imputing missing attribute values of medical records. This is done by clustering medical records which were free from missing values. The records with missing values were separate from dataset. The multi-dimensional medical records are transformed to single dimension. In future, the objective is to see the possibility of other clustering procedures and new approaches to impute missing values. The present approach may be extended to perform classification and prediction without the need for adopting separate procedures to achieve the required objectives. This method is first of its kind which may be used to perform missing values imputation, classification, and disease prediction in a single stretch. A simple common sense shows the importance of the approach carried out and may be extended to any other domain of interest by researchers.

X. Conclusions and Scope for Future Research

In the present research, we address the first challenge of handling missing values in medical datasets. We also address how the dimensionality reduction of medical datasets may be achieved in a simple approach. We come up with a new approach of finding missing values in datasets not addressed in the literature by aiming at a single dimension. The approach followed does not miss any attribute information while carrying out dimensionality reduction which is the importance of this approach. The proposed approach of imputing missing values in medical records is feasible for both categorical and numerical attributes as discussed in case study. However, suitable normalization techniques must be applied, if required for some datasets after extensive study of the datasets. In this paper, we also extend imputation approach also for prediction and classification of unknown medical records for predicting disease levels or symptoms through soft computing techniques. The approach overcomes ambiguity which is otherwise possible if dimensionality reduction is not carried properly.

References

[1] Zhang, S, Zhenxing Qin, Ling C.X, Sheng S, " "Missing is useful": missing values in cost-sensitive decision trees,", IEEE Transactions on Knowledge and Data Engineering, vol.17, no.12, pp.1689-1693, 2005.

[2] Zhang, C,Yongsong Qin, Xiaofeng Zhu, Jilian Zhang, and Zhang,S, "Clustering-based Missing Value Imputation for Data Preprocessing," in , 2006 IEEE International Conference on Industrial Informatics, pp.1081-1086, 2006.

[3] Wang, Ling, Fu Dongmei, Li Qing, Mu Zhichun, "Modelling method with missing values based on clustering and support vector regression," , Journal of Systems Engineering and Electronics , vol.21, no.1, pp.142-147, 2010.

[4] Kirkpatrick B, Stevens K, " Perfect Phylogeny Problems with Missing Values," IEEE/ACM Transactions on Computational Biology and Bioinformatics,Vol.11,No.5,pp.928-941,2014.

[5] Xiaofeng Zhu, Zhang S, Zhi Jin, Zili Zhang, and Zhuoming Xu, "Missing Value Estimation for Mixed-Attribute Data Sets", IEEE Transactions on Knowledge and Data Engineering, Vol.23, No.1, pp.110-121, 2011.

[6] Farhangfar A, Kurgan L.A, Pedrycz ,"A Novel Framework for Imputation of Missing Values in Databases," in Part A: Systems and Humans, IEEE Transactions on Systems, Man and Cybernetics, Vol.37, No.5,pp.692-709, 2007.

[7] Miew Keen Choong,Charbit M, Hong Yan, "Autoregressive-Model-Based Missing Value Estimation for DNA Microarray Time Series Data,",IEEE Transactions on Information Technology in Biomedicine,Vol.13, No.1,pp.131-137, 2009.

[8] Qiang Yang, Ling C, Xiaoyong Chai, and Rong Pan, "Test-cost sensitive classification on data with missing values," in IEEE Transactions on Knowledge and Data Engineering, Vol.18, No.5, pp.626-638, 2006.

[9] G. Madhu, “A Non-Parametric Discretization Based Imputation Algorithm for Continuous Attributes with Missing Data Values”, International Journal of Information Processing, Volume 8, No.1, pp.64-72, 2014.

[10] Sreehari Rao, NareshKumar, ”A New Intelligence-Based Approach for Computer-Aided Diagnosis of Dengue Fever, “ , IEEE Transactions on Information Technology in Biomedicine, Vol.16, Issue 1, pp.112 – 118, 2012.

[11] V. Sree Hari Rao and M.NareshKumar, “Novel Approaches for Predicting Risk Factors of Atherosclerosis”, IEEE Journal of Biomedical and Health Informatics, Vol.17,Issue1 pp. 183 – 189,2013

[12] G.Madhu, “A novel index measure imputation algorithm for missing data values: A machine learning approach”, IEEE International Conference on Computational Intelligence & Computing Research, pp.1-7,2012.

[13] G.Madhu,” A Novel Discretization Method for Continuous Attributes: A Machine Learning Approach”, International Journal of Data Mining and Emerging Technologies, pp.34-43, Vol.4, No.1, 2014.

[14] G.Madhu,“ Improve the Classifier Accuracy for Continuous Attributes in Biomedical Datasets Using a New Discretization Method”, Journal Procedia Computer Science,Vol 31, pp.671-79, 2014.

[15] Atif Khan, John A. Doucette, and Robin Cohen, ” Validation of an ontological medical decision support system for patient treatment using a repository of patient data: Insights into the value of machine learning”, ACM Trans. Intell. Syst. Technol,Vol.4,No.4,Article 68, 31 pages,2013.

[16] Jau-Huei Lin and Peter J. Haug,”Exploiting missing clinical data in Bayesian network modeling for predicting medical problems”, Journal of Biomedical Informatics, Vol.41, Issue 1, pp.1-14, 2008.

[17] Karla L. Caballero Barajas and Ram Akella,” Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach”, In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15), pp..69-78, 2015.

[18] Zhenxing Qin, Shichao Zhang, and Chengqi Zhang. 2006. Missing or absent? A Question in Cost-sensitive Decision Tree. In Proceedings of the 2006 conference on Advances in Intelligent IT: Active Media Technology, Yuefeng Li, Mark Looi, and Ning Zhong (Eds.). IOS Press, pp.118-125,2006.

[19] Shobeir Fakhraei, Hamid Soltanian-Zadeh, Farshad Fotouhi, and Kost Elisevich,“ Effect of classifiers in consensus feature ranking for biomedical datasets”, In Proceedings of the ACM fourth international workshop on Data and text mining in biomedical informatics, DTMBIO '10,pp.67-68, 2010.

[20] Wei-Chao Lin, Shih-Wen Ke, Chih-Fong Tsai, “CANN: An intrusion detection system based on combining cluster centers and nearest neighbors”, Knowledge-Based Systems, Volume 78, pp.13-21,2015.

[21] Aljawarneh, S., Shargabi, B., & Rashaideh, H. (2013). Gene classification: A review. Proceedings of IEEE ICIT.

A study of Adopting Cloud Computing from Enterprise Perspective using Delone and Mclean IS Success Model

Bassam Al-Shargabi, Omar Sabri
Isra University, Amman-Jordan

Abstract: Nowadays, Cloud Computing is the new promising technology that enable sharing resources between different enterprises through Internet in an on-demand manner. Many enterprises are moving toward adopting cloud computing services to gain the benefits of cost reduction of such services. Thus , many enterprises are facing greater obstacles for adapting this new technology, In this paper, the DeLone and McLean successes model is used to assess and evaluate some components that need to be considered by an enterprise when making the decision of adopting cloud computing. The enterprise will be able to identify its weakness and strength for each factor, and then build and prepare plan that can help them to make appropriate decision toward a successful adoption of Cloud Computing.

Keywords: Cloud Computing; Information Technology; Software as a Service; Infrastructure as a Service; Platform as a Service

1. Introduction

Cloud Computing can be presented as a model for hosting and delivering services (software, hardware, platforms) over the internet to a client. Cloud Computing is the next stage in the escalation of the internet although it has different meaning with the internet [1, 2]. It is technological trend that play the role of hosting and delivering services and enable sharing resources between users through Internet in an on-demand manner.

Cloud computing introduces a services to different type of enterprises such as governments, universities, industries, and also small to medium enterprises [3, 4].The most significant impact of cloud computing or technology might appear from cost savings that lead to an increased competitiveness in IT services available to a wide range of enterprises whether its public or private, along with the introducing of new IT services caused by the cloud computing. Because of demand aggregation, bulk purchasing of power and hardware, and reduced per-unit labour costs cloud providers can make a considerable savings on their running costs, and pass these on to their customers or cloud consumers. Businesses can benefit from cloud technologies in area of IT provision, through using equipment better, being more flexible, being faster, and having less capital expenditure. For consumers, cloud technologies are making information and online content more accessible and more interactive[3].

Cloud computing is intelligent to business owners as it reduces the cost of hardware and software resources. The services of cloud computing reduces the costs of computing and communication which increases the interest to companies around the world [2, 3].Many organization and enterprise they believe that cloud computing may offer feasible alternative model that may reduce costs and complexity by increasing operational efficiency However, many organizations are finding greater obstacles for adopting to this new technology, obstacles such as security , privacy, reliability, and economic return value [21]. Cloud Computing is not the ultimate solution for every organization, Therefore, each organization must analyze its existing IT infrastructure and asses their problems with its current solutions, and what are the real benefits of moving toward Cloud Computing. This paper introduces an evolution model for assessing the advantages and the disadvantages of cloud computing on different components depending on DeLone and McLean successes model.

The paper is organized as the following: we give the definitions of Cloud Computing, services, categories, participants, advantages, and limitations of Cloud Computing. DeLone and McLean successes model is introduced to identify the weakness and strength for each factor where the organization can make plan and decision toward a successful adoption of Cloud Computing

2. Cloud Computing Concepts

There are many definitions for cloud computing in the literature. some of the authors do no not consider cloud computing as new technology, however, as a combination of many already exist technologies that make the cloud a more compelling solution [5, 6]. These technologies such as: broadband Internet connection; fast and low cost servers; and advanced processors, virtualization technology, and disk storage are developed and combined to create a technical environment for cloud computing.

Cloud also can be seen as a parallel and distributed computing systems consisting of a collection of inter-connected computers that are dynamically provisioned and presented as one or by more joint computing resources based on service-level agreements (SLA) established through negotiation between the service provider and consumers [7,8].

Cloud computing can be seen as a model for allowing ubiquitous, convenient, on-demand network access through a shared pool of configurable computing resources, or as a compilation of hardware, interfaces, networks, storage, and services, that allow the prompt delivery of software and services over the Internet upon user demand [1, 5, 7].

The national institute of standard and technology of America defined Cloud Computing as the following [5, 9]:“Cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.”

Cloud computing improves major trends in information technology (IT efficiency, Business agility) For IT efficiency, cloud computing efficiently utilizes the power of modern computers through the scalability of software and hardware resources. Also Business agility is prospered because of the fast deployment of real time applications into interactive mobile devices, and improving the intensive computing of business analyses [5, 6].

3. Cloud Computing Services

There are three service model for cloud computing services [3,10, 11], each model describes how these services are available to client, those models includes IaaS (Infrastructure as a Service), PaaS (Platform as a Service), SaaS (Software as a Service)

IaaS (Infrastructure as a Service): which is the most fundamental model in cloud computing that provides infrastructure component for clients through integrating the hardware components such as I/O devices, memory, and storage into a virtual resources pool. This will reduce hardware cost since the user will pay only for his need.

PaaS (Platform as a Service): It is platform server distribution where users customize and develop their own application and transfer to other customers through their server and Internet. The platforms offered include development tools, configuration management, and deployment platforms.

SaaS (Software as a Service): is subscribing software services from the manufacturer through Internet. The Provider supply software such as spreadsheet tools, or CRM services then charge according to the duration of use and the quantity of the software. A distinctive properties of SaaS exploitation does not involve any hardware and can be handled through the existing Internet access only. Occasionally, enterprise might need to change their firewall rules and settings to permit the SaaS application to run efficiently.

3. Cloud Computing Categories

Cloud computing could be classified into four categories: (The Public Cloud, The Private Cloud, Community cloud, and The Hybrid Cloud Computing) [1, 2, 3, 7, 12, 13,14]

The Public Cloud: is set of computing resources that is provided by third party that supports on demand computing resources such as applications and web-services are provided over the Internet from an off-site third-party service provider. Sometimes most enterprises they can benefit from this type in term of saving cost for buying or maintaining their own IT infrastructure.

The Private Cloud is Cloud Computing over private networks exclusively used by everyone in the organization, but not for others out of the organization. Clients have full control of data, services and applications. Mainly enterprises can benefit from building their own private clouds, according to their existing IT infrastructure . This can be considered as the first step in a any enterprise to move from existing IT systems towards public cloud services. at the same time as the private cloud does not generally provide as great cost savings as a public cloud, it might be appropriate for enterprises with sensitive data they do not want to send outside their own systems.

.

Community cloud Community cloud are prepared to make services and resources for specific community of users or organizations that follow the same mission, goals, security and policy among other requirements. The size and number of the organisations, amongst other factors, determines the extent of demand, and hence cost savings they can obtain.

The Hybrid Cloud Computing it is a mix between a variety of public and private Clouds where some enterprises can do some computation and storage in the public cloud and some on in their private cloud that might contain their sensitive data, which allows more sensitive data to be processed in-house while less sensitive data goes to cheaper public cloud servers.

4. The Participants in Cloud Computing

End user is client or customer who does not know or care about the technology in use. Cloud provider is considered as the data center for small to mid size businesses, while IT enterprises manage the cloud resources for large organizations. Business management has a responsibility for data or services control in the Cloud. Cloud service provider has a responsibility to sustain the IT resources.

5. Cloud Computing Benefits for Enterprises

the enterprises can have several benefits and advantages when they decide to adapt Cloud Computing [16, 17]. these advantages are listed below:

Reducing cost: enterprises don not need a high performance computer to run their application since all their application are accessible through the cloud. Since applications run in the cloud, a desktop computer does not need the processing power or hard disk space demanded by traditional desktop software, or even a DVD drive since distributed applications are accessed by browsers, which reduce the cost for owners and users.

Infrastructure liability : organization can reduce the risks of any liability issue regarding the loss or damage of thier IT infrastructure its applications

Accessibility: applications are accessed autonomously by users from any location, since cloud can be accessed from anywhere..

Enhance storage area: distributed process, allows more storage than centralized storage.

Flexibility: Cloud Computing provides a download free zone where software can be upgraded, administrated, installed by its own.

Mobility: The Cloud can be connected from any location.

Ease of sharing: one of the important features of Cloud Computing where resources, information, and hardware are shared for instant delivery.

Data safety: The shared files are safe unless the hard drive is stolen.

Availability: There are numerous copies which are distributed in different servers and can be retrieved on demand

Copyright: Licensing and authorization are reserved.

Portability: software is transferred from one platform to another.

On-demand: prompt delivery of software and services

Scalability: users can access huge pool of virtual resources

5. Enterprises Obstacles of Cloud Computing

As there are advantages for cloud computing there are disadvantages or some limitations should be considered for enterprises in order to be eased or accommodated when they decide to move toward cloud computing[1, 5, 8, 6], such limitatition and other issue discussed below:

Security: in IT fields most decisions are taken based on security risk. Information executives consider security as the first important concern in Cloud Computing. Hence, all security levels such as data, application, and network should be confident. Enterprises are worried who control or see their data and they would feel safe since their data are secure in their internal IT infrastructure as compared to the cloud. Therefore, the ability of cloud computing to sufficiently address privacy policy and data integrity has been called into question, more detailed information about this issue is presented in section 6 .

Interoperability: In Cloud Computing, each Cloud has its own platform, programming syntax, and data storage. Many companies have made large steps forward standardizing their systems, methods and interfaces through implementation of ERPs.

Consistency assurance: Due to data and code replication over the Cloud, it is very critical to keep consistency in data.

Reliability: Cloud services may fail and users may lose data. Organizations must make their applications reliable and available. Emergency plans are required in the case of failure or outages and recovery plans are required for disastrous failure. Of course extra costs may be associated with the necessary levels of reliability; however, it is feasible in contrast to the cost of a failure.

Ambiguity of Cost: cloud provider vary in their price model , thus , which make it very difficult to estimate financial return of using cloud computing

Liability and political issue: In world of Cloud computing , where the data resides for enterprises is major problem, or where processing takes place , who is accessing to data. Regarding those issue , different privacy rules and regulations must be imposed. Therefore, different rules and regulations are to be applied depending on the location of data or processing. There is an eminent need for global rules and regulation to govern the use of Cloud Computing in order to resolve the different rules and regulation around the world . Nowadays, the key global technological and political powers are enforcing regulations that may have a negative impact enterprises to adopt Cloud Computing[18].

6. Security in Cloud Computing

Since the data security in the cloud is a major problem or concern for enterprise to adopt cloud computing, in this section a detailed discussion of major important security issues facing cloud computing [1, 12].

The risk of hacking stored data or violence of transmitting data over the Clouds. It is difficult for cloud customer as a data controller to insure that data is handled in a lawful way. Some cloud providers give information on their data handling practices.

Sharing resources: in cloud computing may cause failure of resource separate mechanisms between tenants.

Lock-in: is the difficulty occurs when the customer needs to migrate from one provider to another. When deleting any cloud computing resources, the deletion may not be completed due to the redundant copies of the data.

Loss of compliance: usually Cloud Provider controls number of issues that may affect the security when using cloud infrastructures such as compliance risks, because industry standard or regulatory requirements may be put at risk by migration to the cloud.

Malicious insider is less likely to occur but the damage that may be caused by it is far larger.

enterprises perceptions about security levels may differ from the actual security offered by the cloud provider.

7. Delone and Mclean IS Success Model to Adopt Cloud Computing

The updated DeLone & McLean[16] success model provides a comprehensive framework for assessing and evaluating IS success. This model were used in developing IS in Various disciplines such as e-commerce, e-learning, meta-analysis, and knowledge transfer and other fields [17].

In this paper the updated Delone and McLean model is adopted to analyze and measure the success factors from a enterprises point view when making the decision of using cloud computing.

The IS success is classified into six major dimensions: Information quality, System quality, Service quality, Use, User satisfaction, and Net benefits as illustrated in figure1.The updated

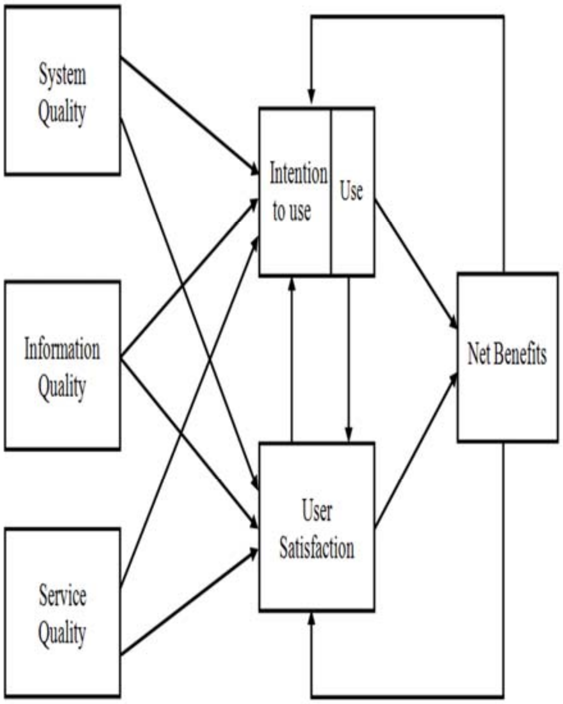


Fig. 1. Updated DeLone and McLean IS Success Model [17]

dimensions of DeLone and McLean IS Success Model and how this model can be used by enterprises to plan for a successful adopting Cloud Computing:

System quality: measuring the technical issues such as the desirable feature of system flexibility, reliability, fast response, and ease of use. In terms of flexibility Cloud Computing download free zone where software can be upgraded, administrated, and installed by its own. Also Cloud Computing is On-demand services. However, in terms of reliability Cloud services may fail and users may lose data so Organizations must have Emergency plans to make their applications reliable and available.

Information quality: measures the semantic success in terms of understandability, accuracy, usability, completeness, and timeliness. Timeliness and Ease of sharing are of the important features of Cloud Computing. Because software is transferred from one platform to another, companies have to make large steps forward standardizing their systems, methods and interfaces through implementation of ERPs.

Service quality: is the quality of services gained to users from IS department. The Availability of Cloud Computing due to the distributed copies in different servers allows the data to be retrieved on demand. However, the availability of Internet access is affected by politics. Also Cloud Computing is scalable where users can access huge pool of virtual resource. Cloud Computing enable reducing cost and enhance accessibility where applications are accessed autonomously by users from any location. Economic objectives in the company should be defined related financial, customer, internal and learning-development then identify the way cloud services can maintain these objectives [5]. In this stage also an organization should take into its consideration all security issues discussed previously.

System use: the amount of how much people gain the capabilities of an IS. Portability is one of the features in Cloud Computing where software is transferred from one platform to another. However, organization should consider the difficulty occurs and performs standards to permit maintenance of the integrity and consistency of its information. Also availability relies on Internet connectivity at customer’s end.

User satisfaction: like the user’s Attitudes, expectations and involvement toward technology. Ease of sharing is one of the important features of Cloud Computing where resources, information, and hardware are shared for instant delivery. This feature enhances user satisfaction.

Net benefits: the contribution of the IS to the success of organizations by measuring productivity improvement, cost reductions, and user interests. Successful Cloud Computing outcomes will result in better organization outcomes.

Eventually all enterprises must take in consideration if the Cloud Computing can serve and support their strategic planning and their vision as the study in [19,20].

8.Conclusion

Despite the advantages resulting from Cloud Computing from the enterprises perspective, each enterprise must evaluate the advantages and disadvantages of adopting the Cloud Computing. This study aimed to apply the updated DeLone and McLean IS success model to measure the success components of Cloud Computing in enterprise perspective.

The framework could be used by enterprises to consider the success components by identifying the weakness and strength points of the components, and then build a preparing plan that can help them to achieve the readiness required towards Cloud Computing success from enterprises perspective.

Acknowledgment

I would like to thank everyone who contributed to the completion of this work.

Reference

[1] Bhatta, D. 2012. Revolution in Information Technology - Cloud Computing. Walailak J. Sci. & Tech., 2012; 9(2) 107–113

[2] Hashemi, S. 2013.Cloud Computing Technology For E-Government Architecture. International Journal in Foundations of Computer Science & Technology (IJFCST), 3(6) .

[3] Civic Consulting, 2012,. Economic and scientific policy cloud computing. Technical report. http://www.europarl.europa.eu. accessed on 5/9/2015

[4] Marawar, T, Kale, S, Araspure, K. 2010. E-Governance ,IEEE Software, 183-186.

[5] M.G. Avram. 2014. Advantages and challenges of adopting cloud computing from an enterprise perspective. The 7th International Conference Interdisciplinarity in Engineering (INTER-ENG 2013). Procedia Technology 12 (2014): 529 – 534

[6] Manolov ,S., Trifonov,R. 2010. Specific Information Security Problems For Cloud Computing E-Government applications, Proceedings of the International Conference on Information Technologies (InfoTech-2010) 16-17 September 2010, Bulgaria

[7] Choudhary,V. and Vithayathil,J.2013. The Impact of Cloud Computing: Should the IT Department Be Organized as a Cost Center or a Profit Center?. Journal of Management Information Systems. 30(2): 67–100.

[8] Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I.2009. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computer Systems,

[9] Budniks, L. and Didenko, K.2014. Factors determining application of cloud computing services in Latvian SMEs. Procedia - Social and Behavioral Sciences 156 (2014) 74 – 77.

[10] Zhang,S. Yan ,H.,and Chen,H.2012. Research on Key Technologies of Cloud Computing. International Conference on Medical Physics and Biomedical Engineering. Physics Procedia 33 (2012) 1791 – 1797.

[11] Stipravietis,P., Zeiris, E., and Ziema, M. 2013. Electronic Service Architecture Model Assessment of Conformity to Cloud Computing Key Features. Technologies of Computer Control.2013 (14):86-95.

[12] Andrikopoulos,V., Binz,T., Leymann,F.,and Strauc,S. 2013. How to adapt applications for the Cloud environment Challenges and solutions in migrating applications to the Clou. Computing (2013) 95:493–535

[13] Rev.B . 2012 . Cloud Computing Benefits, risks and recommendations for information security. Enisa European Network and Security Information Agency. December 2012.

[14] Gusev,M., Ristov,S., Armenski,G., Velkoski,G. and Bozinoski,K. 2013.E-Assessment Cloud Solution: Architecture, Organization And Cost Model. International Journal of Emerging Technologies in Learning (iJET).8(2):55-65.

[15] Alexandru. I.,C.2012. Cloud Computing Based Information Systems -Present And Future. The USV Annals of Economics and Public Administration. 12(16).

[16] William, H. DeLone and Ephraim R. McLean.2003. The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. Journal of Management Information Systems. 19(4):9–30.

[17] Halonen, Raija. 2011. Reflecting With The Delone & Mclean Model. International Workshop on Practice Research in Helsinki.

[18] Pragaladan, R , Suganthi ,P.2014 A Study on Challenges of Cloud Computing in Enterprise Perspective. International Journal of Advanced Research in Computer and Communication Engineering3.(7):7405-7410

[19] Alharbi,F, Atkins, A.Stainer,C.2015. Strategic Framework for Cloud Computing Decision-Making in Healthcare Sector in Saudi Arabia. The Seventh International Conference on eHealth, Telemedicine, and Social Medicine.25-27 ebruary,2015. Lisbon,Portugal: 138-144.

[20] Almabhouh,A.2015.Opportunities of Adopting Cloud Computing in Palestinian Industries. International Journal of Computer and Information Technology. 4(1):103-109.

[21] Aljawarneh, S., Alkhateeb, F., & Al Maghayreh, E. (2010). A semantic data validation service for web applications. Journal of Theoretical and Applied Electronic Commerce Research, 5(1), 39–55.

Clustering and Classification of Text Documents Using Improved Similarity Measure

G.SureshReddy[1], T.V.Rajinikanth[2], A.AnandaRao[3]

[1]Department of Information Technology, VNR VJIET, Hyderabad, India
[3]Department of Computer Science and Engineering, JNTU University, Anantapur, India
[2]Department of Computer Science and Engineering, SNIST, Hyderabad, India

Abstract: Dimensionality reduction is very challenging and important in text mining. We need to know which features be retained what to be and It helps in reducing the processing overhead when performing text classification and text clustering. Another concern in text clustering and text classification is the similarity measure which we choose to find the similarity degree between any two text documents. In this paper, we work towards text clustering and text classification by addressing dimensionality reduction using SVD followed by the use of the proposed similarity measure which is an improved version of our previous measure [25, 31]. This proposed measure is used for supervised and un-supervised learning. The proposed distance measure overcomes the disadvantages of the existing measures [10].

Keywords: Feature Vector, Similarity, Feature Set, Commonality

1. Introduction

Text mining may be defined as the field of research which aims at discovering; retrieving the hidden and useful knowledge by carrying out automated analysis of freely available text information and is one of the research fields evolving rapidly from its parent research field information retrieval [1]. Text mining involves various approaches such as extracting text information, identifying and summarizing text, text categorization and clustering. Text Information may be available either in structured form or unstructured form.

One of the widely studied data mining algorithms in the text domain is the text clustering. Text clustering may be viewed as an unsupervised learning approach which essentially aims at grouping all the text files which are of similar nature into one category thus separating dissimilar content in to the other groups. In contrast to the text clustering approach, the process of text classification is a supervised learning technique with the class labels known well before. In this paper, we limit our work to text clustering and classification. Clustering is a NP-hard problem. One common challenge for clustering is the curse of dimensionality which makes clustering a complex task. The second challenge for text clustering and classification approaches is the sparseness of word distribution. The sparseness of features makes the classification or clustering processes in accurate, in efficient and thus becoming complex to judge the result.

The third challenge is deciding the feature size of the dataset. This is because the features which are relevant may be eliminated in the process of noise elimination. Also deciding on the number of clusters possible is also a complex and debatable.

In this paper, we carry out the dimensionality reduction at two stages. The first stage of dimensionality reduction takes in to the consideration elimination of stop words, stemmed words, followed by computation of tf-idf. The second stage of dimensionality reduction is by the use of singular valued decomposition approach. This is followed by the use of proposed improved similarity measure w.r.t similarity measure [25].
The proposed measure is applied to supervised learning process and also for

the un-supervised learning process. Section-2 of this paper discusses the related works. Section-3 introduces the proposed measure and dimensionality reduction [31] .Section-4 deals case study. Section-5 and section-6 text clustering and classification. Section-7 concludes the paper.

2. Related Work

Text mining spans through various areas and has its applications including recommendation systems, tutoring, web mining, healthcare and medical information systems, marketing, predicting, and telecommunications to specify a few among many applications[1]. The authors [2], study and propose various criteria for text mining. These criteria may be used to evaluate the effectiveness of text mining techniques used. This makes the user to choose one among the several available text mining techniques. In [3], the authors use the concept of text item pruning and text enhancing and compare the rank of words with the tf-idf method. Their work also includes studying the importance and extending the use of association rules in the text classification.

Association rule mining is playing an important role in text mining and is also widely studied, used and applied by the researchers in text mining community. In [4] authors, discuss the importance of text mining in the predicting and analyzing the market statistics. In short, they perform a systematic survey on the applicability of text mining in market research. In [5], the authors work towards finding the negative association rules. Earlier in the past decade, the data mining researchers and market analysts were only interested in finding the dominant positive association rules. In the recent years, significant research is carried out towards finding the set of all possible negative association rules. The major problem with finding negative association rules is the large number of rules which are generated as a result of mining. The negative association rules have important applications in medical data mining, health informatics and predicting the negative behavior of market statistics. In [6], the authors use the approach of first finding the frequent items and then using these computed frequent items to perform text clustering. They use the method called “maximum capturing”. With the vast amount of information generating in the recent years, many researchers started coming out with the extensive study and defining various data mining algorithms for finding association rules, obtaining frequent items or item sets, retrieving closed frequent patterns, finding sequential patterns of user interest[7]. All these algorithms are not suitable for their use in the field of text mining because of their computational and space complexities. The suitability of these techniques in text mining must be studied in detail and then applied accordingly. One of the important challenges in text mining is handling the problems of misinterpretation and less frequency. An extensive survey on dimensionality reduction techniques is carried in [8]. The authors discuss the method of principal factor analysis, maximum likely hood factor analysis and PCA (principal component analysis). A fuzzy approach for clustering features and text classification which involves soft and hard clustering approaches is discussed in [12]. An improved similarity measure overcoming the dis-advantages of conventional similarity measures is discussed in [10]; their work also involves clustering and classification of text documents. In [11], the concept of support vector machines, SVM is used for document clustering. The other significant findings and research works in the area of text mining include work by the researchers [16-23]. In the present work, our idea is to design a similarity measure overcoming dis-advantages in Euclidean, Cosine, Jaccard distance measures [10]. The proposed measure considers distribution of features of the global feature set.

3. Problem Definition

We divide the problem definition into 3 stages

1. Designing Feature Function
2. Represent corresponding Feature vector
3. Design of similarity function
4. Validation similarity measure

.

The objective is to design a similarity measure which can estimate similarity between any two chosen text files. The measure designed is based on the presence-absence of a feature being considered. We consider three possibilities to design proposed measure

1. ith feature is present in both text files
2. ith feature is absent in both text files
3. ith feature is present in one of text files

3.1 Feature Function and Vector

We denote function $\boldsymbol{f}_{\mathbf{c}}$**<w$^{\mathbf{(1m)}}$,w$^{\mathbf{(2m)}}$ >** defined as shown in the Table.1. Here, **w$^{\mathbf{(1m)}}$** and **w$^{\mathbf{(2m)}}$** indicate the presence or absence of the mth feature in ith and jth text files indicated by f^{i} and f^{j} respectively.

The presence of feature in text file is denoted by a value 1 and its absence by a value 0. The feature function, f_c< w^{1m},w^{2m} > evaluates to any one of the values 0, 1 or -1.

Table.1 Feature Function $\boldsymbol{f}_{\mathbf{c}}$< **w$^{\mathbf{(1m)}}$, w$^{\mathbf{(2m)}}$** >

w$_{1m}$	w$_{2m}$	$\boldsymbol{f}_{\mathbf{c}}$< **w$^{\mathbf{(1m)}}$, w$^{\mathbf{(2m)}}$** >
absence (0)	absence (0)	-1
absence (0)	presence (1)	1
presence (1)	absence (0)	1
presence (1)	presence (1)	0

We represent text files F$_i$ and F$_j$ as

F^i = { w^{i1}, w^{i2}, w^{i3}, w^{i4}, w^{i5}... w^{im} } and
F^j = { w^{j1}, w^{j2}, w^{j3}, w^{j4}, w^{j5}... w^{jm} }.

The notations w^{im}, w^{jm} represents presence or absence of mth feature in ith and jth text files respectively. The presence of the feature w^{im} is denoted by 1 and its absence by 0. Now, we define generalized feature vector expressed as a function of feature function defined in table.1 above.

Let F^1 and F^2 be any two text files, then the feature vector for these two files is denoted using notation FV^{12} and formally represented as Feature-vector [F^1, F^2],

FV$_{12}$ = [f_c< ^{w11} ,^{w21} >, f_c < w^{12} ,w^{22} > .... f_c < w^{1m} , w^{2m}>].

3.2 Proposed Measure

The similarity measure is given by the equation 1 below

$$Sim = \frac{(N_{avg} + \lambda)}{(\lambda + 1)} \qquad (1)$$

where

$$N_{avg} = \frac{\sum_{k=1}^{k=m} N_{nr}(F^{ik}, F^{jk})}{\sum_{k=1}^{k=m} N_{dr}(F^{ik}, F^{jk})} \qquad (2)$$

and

$$N_{nr}(F^{ik}, F^{jk}) = \begin{cases} [1 - e^{-(\frac{1-f_c<w^{ik},w^{jk}>}{\sigma k})^2}] ; f_c < w^{ik}, w^{jk} >= 0 \\ -\lambda e^{-(\frac{1-f_c<w^{ik},w^{jk}>}{\sigma k})^2} \quad ; f_c < w^{ik}, w^{jk} >= 1 \\ 0 \quad ; f_c < w^{ik}, w^{jk} >= -1 \end{cases}$$

(3)

$$N_{dr}(F^{ik}, F^{jk}) = \begin{cases} 1 ; f_c < w^{ik}, w^{jk} > \neq -1 \\ 0 ; f_c < w^{ik}, w^{jk} >= -1 \end{cases}$$

(4)

N_{avg} is fixed to – λ incase N_{nr} and N_{dr} are both evaluated to 0. Similarity value is evaluated to a value between 0 and 1. A value 0 indicates similarity between two text files is minimum and a 1 indicates similarity between two text files is maximum. λ is fit to 1.

A threshold value δ may be defined by the user to select all the text files whose similarity is above the value of S. The ratio of $N_{nr}(F^{ik}, F^{jk})$ and $N_{dr}(F^{ik}, F^{jk})$

denotes the average contribution of features of text files.

The proposed measure is validated in the sections below.

3.3 Properties of Proposed Measure

The proposed measure is applicable for both frequency and binary representations of document vectors.

Property 1: It satisfies symmetric property. i.e distance between f^i and f^j .

Property 2: The standard deviation of each feature from the global feature set is considered for its contribution to find similarity between text documents.

Property 3: The degree of similarity decreases as the number of presence-absence feature increases and vice versa. i.e the similarity degree and presence-absence feature are inversely proportional.

Property 4: The degree of similarity increases when both the documents have all features and further increases when these features individual distribution is widely spread across all the text documents.

Property 5: Two text documents are least similar when two documents do not have at least one feature from global feature vector. In worst case, the similarity value is 0.

Property 6: Two text documents have maximum similar when two documents have all features. In the best case, the maximum similarity value is 1.

Property 7: Two text documents have average similarity when the document vectors are combinations both features present and also the presence-absence features.

Property 8: There is finite lower bound and upper bound for the defined similarity measure. For example, Euclidean measure has no finite upper bound and can be even infinite.

Property 9: The presence or absence of a feature is more important than the frequency count features to estimate the similarity between two text documents.

3.4 Remarks

Remark-1: (Property 3, 7, 9)

Let two documents be denoted by F^{pk} and F^{qk} representing k^th feature in the text documents F^p and F^q respectively. A value $F^{pk}= 0$ indicates absence of the feature and $F^{pk}= 1$ indicates presence of feature, k.

Consider the Case-1 and Case-2 as shown below

Case-1: $F^{pr} =1$ and $F^{qr} =1$

Case-2: $F^{pr} =1$ and $F^{qr} =0$
(or)
$F^{pr} =0$ and $F^{qr} =1$

A simple common sense, indicates the similarity value computed for the case-1 must be greater than case-2. Here $N_{dr}\left(F^{ik},F^{jk}\right)$ remains same for both the situations and this value is denoted as y. However, $N_{nr}\left(F^{ik},F^{jk}\right)$ remains different for both situations. Let x denotes

$$\sum_{k=1,k\neq r}^{k=m} N_{nr}\left(F^{ik},F^{jk}\right)$$

Case-1: $$N_{avg} = \frac{x + [1-e^{-(\frac{1}{\sigma_k})^2}]}{y} \qquad (5)$$

Case-2: $$N_{avg} = \frac{x - \lambda e^{-(\frac{1}{\sigma_k})^2}}{y} \qquad (6)$$

Since the value obtained by $(x + e^{-(\frac{1}{\sigma_k})^2}) > (x - e^{-(\frac{1}{\sigma_k})^2})$ is obviously greater. Hence, the similarity value for the first case is greater than the second case.

Remark-2: (Property 5, 8)
The similarity value defined by G_{SIM} is least when both text documents do not contain non-zero values indicating all features are absent in both the documents. Consider the Case-3

Case-3: $F_{pr} = 0$ and $F_{qr} = 0$.

Here, $\beta(F_{ik}, F_{jk}) = 0$, $\alpha(F_{ik}, F_{jk}) = 0$.

Then, $N_{avg} = \frac{0}{0}$ (7)

In such a case, we return -1 as the value of N_{avg}. With the value of λ fixed to 1, this gives the similarity value as

$$G_{SIM} = \frac{(\lambda - 1)}{(1 + \lambda)} = 0 \quad (8)$$

Remark-3: (Property 4, 6, 8)
The similarity value defined by G_{SIM} is maximum when both the text documents contain the non-zero values indicating all features are present in both the documents. Consider the Case-4

Case-4: $F_{pr} = 1$ and $F_{qr} = 1$

Here, $\beta(F_{ik}, F_{jk}) = 1$, $\alpha(F_{ik}, F_{jk}) = 1$

Then,$N_{avg} = \frac{1}{1} = 1$ (9)
In such a case, we return -1 as the value of N_{avg}. This gives the similarity value as

$$G_{SIM} = \frac{(1 + N_{AVG})}{(1 + 1)}$$
$$= 1 \quad (10)$$

Remark-4: (Property, 9)
Consider the situation, where $F_{ik} = 0$ *and* $F_{jk} = 6$, we know that the documents F_i *and* F_j are not similar with respect to kth feature.

Similarly, $F_{ik} = 6$ *and* $F_{jk} = 14$ indicates some similarity between the documents. In this case the difference remains same. In such a case, the frequency of features loses its importance.

From this, we can conclude that kth feature has more importance in the presence-absence situation than when both the text documents contain the kth feature.

Remark-4: (Property 1, 2)
Since the similarity value is based on standard deviation of kth feature, but on the order of files or any other parameter, in our case the similarity measure is symmetric.

3.5 Analysis of Proposed Measure

3.5.1 Best Case

For best case situation, each feature is present in text files being considered. For sake of analysis, we assume two files as given below.

Let

$f_1 = \{1, 1, 1, 1, 1.....m\}$ and

$f_2 = \{1, 1, 1, 1, 1........m\}$

The computation values of N_{nr} and N_{dr} values of are shown below for each feature in the feature vector

For i=1, j=2, k=1 : $N_{nr}\ (F^{11}, F^{21}) = 1$

For i=1, j=2, k=2 : $N_{nr}\ (F^{12}, F^{22}) = 1$

For i=1, j=2, k=3 : $N_{nr}\ (F^{13}, F^{23}) = 1$

For i=1, j=2, k=4 : $N_{nr}\ (F^{14}, F^{24}) = 1$

For i=1, j=2, k=5 : $N_{nr}\ (F^{15}, F^{25}) = 1$

………………
………………………
…………………………………

For i=1, j=2, k=m : $N_{nr}\ (F^{1m}, F^{2m}) = 1$

k= m

$\sum N_{nr}(F^{ik}, F^{jk})$ = 1+1+1+1+1…m = m
k= 1

For i=1, j=2, k=1 : N_{dr} (F^{11}, F^{21}) = 1
For i=1, j=2, k=2 : N_{dr} (F^{12}, F^{22}) = 1
For i=1, j=2, k=3 : N_{dr} (F^{13}, F^{23}) = 1
For i=1, j=2, k=4 : N_{dr} (F^{14}, F^{24}) = 1
For i=1, j=2, k=5 : N_{dr} (F^{15}, F^{25}) = 1
………………
……………………….
……………………………….

For i=1, j=2, k=m: N_{dr} (F^{1m}, F^{2m}) = 1

k= m
$\sum N_{dr}$ (F^{ik}, F^{jk}) = 1+1+1+1+1…m = m
k= 1

This gives the value of N_{avg} as

$$N_{avg} = \frac{\sum_{k=1}^{k=m} N_{nr}(F^{ik}, F^{jk})}{\sum_{k=1}^{k=m} N_{dr}(F^{ik}, F^{jk})}$$

$$= \frac{m}{m} = 1$$

The similarity value for best case is hence evaluated to

$$Sim = \frac{(N_{avg}+\lambda)}{(\lambda+1)}$$

$$= \frac{(1+\lambda)}{(\lambda+1)}$$

$$= 1$$

The function $N_{dr}(F^{ik}, F^{jk})$ is introduced to keep track of the set of all the features of the global feature set which contribute to the clustering process and hence all such features must be considered.

It is finally reduced to store the total count of features or to store the distribution count.

3.5.2 Worst Case

For worst case situation, each feature is not present in text files being considered. For sake of analysis, we assume two files as given below.

Let

f_1 = {0, 0, 0, 0, 0…..m} and
f_2 = {0, 0, 0, 0, 0……..m}

The computation values of N_{nr} and N_{dr} values of are shown below for each feature in the feature vector

For i=1, j=2, k=1 : N_{nr} (F^{11}, F^{21}) = 0

For i=1, j=2, k=2: N_{nr} (F^{12}, F^{22}) = 0

For i=1, j=2, k=3: N_{nr} (F^{13}, F^{23}) = 0

For i=1, j=2, k=4: N_{nr} (F^{14}, F^{24}) = 0

For i=1, j=2, k=5: N_{nr} (F^{15}, F^{25}) = 0
………………
……………………….
……………………………….

For i=1, j=2, k=m: N_{nr} (F^{1m}, F^{2m}) = 0

k= m
$\sum N_{nr}(F^{ik}, F^{jk})$ = 0+0+0+0+0…m = 0
k= 1

For i=1, j=2, k=1 : N_{dr} (F^{11}, F^{21}) = 0

For i=1, j=2, k=2 : N_{dr} (F^{12}, F^{22}) = 0

For i=1, j=2, k=3 : N_{dr} (F^{13}, F^{23}) = 0

For i=1, j=2, k=4 : N_{dr} (F^{14}, F^{24}) = 0

For i=1, j=2, k=5 : N_{dr} (F^{15}, F^{25}) = 0
………………
……………………….
……………………………….

For i=1, j=2, k=m : N_{dr} (F^{1m}, F^{2m}) = 0

k= m
$\sum\ N_{dr}\ (F^{ik}, F^{jk})$ = 0+0+0+0+0…m = 0
k= 1

In this case, since both the numerator and denominator of $\mathrm{N_{avg}}$ evaluate to 0 and the measure must return a value -1 to indicate the abnormal situation as stated earlier.

This gives the value of $\mathrm{N_{avg}}$ as

$$N_{avg} = \frac{\sum_{k=1}^{k=m} N_{nr}(F^{ik}, F^{jk})}{\sum_{k=1}^{k=m} N_{dr}(F^{ik}, F^{jk})}$$

$$= \quad -1$$

The similarity value for worst case is hence evaluated to 0, since lambda, λ is fixed to -1

$$Sim = \frac{(N_{avg}+\lambda)}{(\lambda+1)}$$

$$= \frac{(-1+\lambda)}{(\lambda+1)}$$

$$= \quad 0$$

3.5.3 Average Case

We divide this situation in to two situations. The first is the worst situation in average case and second includes average case in general.

For average case situation, we have presence-absence combination of features For sake of analysis; we assume two files as given below.

Let

$\mathrm{f_1}$ = {0, 1, 0, 1, 1…..m} and

$\mathrm{f_2}$ = {1, 0, 1, 0, 0……..m}

The computation values of N_{nr} and N_{dr} values of are shown below for each feature in the feature vector

For i=1, j=2, k=1 : $N_{nr}\ (F^{11}, F^{21})\ = -\lambda$

For i=1, j=2, k=2 : $N_{nr}\ (F^{12}, F^{22})\ = -\lambda$

For i=1, j=2, k=3 : $N_{nr}\ (F^{13}, F^{23})\ = -\lambda$

For i=1,j=2, k=4 : $N_{nr}\ (F^{14}, F^{24})\ = -\lambda$

For i=1, j=2, k=5 : $N_{nr}\ (F^{15}, F^{25})\ = -\lambda$
………………
………………………
…………………………………
For i=1, j=2, k=m : $N_{nr}\ (F^{1m}, F^{2m}) = -\lambda$

k= m
$\sum\ N_{nr}(F^{ik}, F^{jk})$
k=1

$= -\lambda\ -\lambda - \lambda$ ……..m times

$= -\lambda * m$

The computation of N_{dr} is shown below

For i=1, j=2, k=1: $N_{dr}\ (F^{11}, F^{21})$ = 1

For i=1, j=2, k=2: $N_{dr}\ (F^{12}, F^{22})$ = 1

For i=1, j=2, k=3: $N_{dr}\ (F^{13}, F^{23})$ = 1

For i=1, j=2, k=4: $N_{dr}\ (F^{14}, F^{24})$ = 1

For i=1, j=2, k=5: $N_{dr}\ (F^{15}, F^{25})$ = 1
………………
………………………
…………………………………
For i=1, j=2, k=m: $N_{dr}\ (F^{1m}, F^{2m})$ = 1

k= m
$\sum\ N_{dr}\ (F^{ik}, F^{jk})$ = 1+1+1+…m times = m
k= 1

In this case, value of $\mathrm{N_{avg}}$ is

$$N_{avg} = \frac{\sum_{k=1}^{k=m} N_{nr}(F^{ik}, F^{jk})}{\sum_{k=1}^{k=m} N_{dr}(F^{ik}, F^{jk})}$$

$$= \quad -\lambda$$

The similarity value for worst case is hence evaluated to 0, since lambda, λ is fixed to -1

$$Sim \;=\; \frac{(N_{avg}+\lambda)}{(\lambda+1)}$$

$$= \frac{(-\lambda+\lambda)}{(\lambda+1)}$$

$$= \frac{(0)}{(\lambda+1)} = 0$$

3.6 Fixing λ value

Case-1: $\lambda \;=\; 0$

Incase $\lambda \;=\; 0$, the similarity measure is same as the N_{avg} .

$$Sim \;=\; \frac{(N_{avg}+\lambda)}{(\lambda+1)}$$

$$= N_{avg}$$

Best Case:

In best case, $N_{avg} = 1$. The similarity function now reduces to

$$Sim \;=\; \frac{(N_{avg}+\lambda)}{(\lambda+1)}$$

This reduces the similarity measure to a value equal to 1 in the best case situation.

$$Sim \;=\; \frac{(1+0)}{(0+1)} = 1$$

Worst Case:

In best case, $N_{avg} = 0$. The similarity function

$$Sim \;=\; \frac{(N_{avg}+\lambda)}{(\lambda+1)}$$

now reduces to

$$Sim \;=\; \frac{(0+0)}{(0+1)} = 0$$

This reduces the similarity measure to a value equal to 0 in the worst case situation.

Average Case:

In average case, $N_{avg} = 0$. The similarity function

$$Sim \;=\; \frac{(N_{avg}+\lambda)}{(\lambda+1)}$$

Now reduces to

$$Sim \;=\; \frac{(-\lambda+\lambda)}{(\lambda+1)} = 0$$

This reduces the similarity measure to a value equal to 0 in the average case situation.

Case-2: $\lambda \;=\; 1$

$$Sim \;=\; \frac{(N_{avg}+\lambda)}{(\lambda+1)}$$

$$= \frac{(N_{avg}+1)}{2}$$

In best case, if N_{avg} = 1, then similarity between text files, Sim = 1.

In worst case, if N_{avg} = -1, then similarity between text files, Sim = 0.

In average case, if N_{avg} = 0, then similarity between text files, Sim = 0.5.

Best Case:

In best case, $N_{avg} = 1$. The similarity function now reduces to

$$Sim = \frac{(N_{avg} + \lambda)}{(\lambda + 1)}$$

This reduces the similarity measure to a value equal to 1 in the best case situation.

$$Sim = \frac{(1 + 1)}{(1 + 1)} = 1$$

Worst Case:

In worst case, $N_{avg} = 0$. The similarity function

$$Sim = \frac{(N_{avg} + \lambda)}{(\lambda + 1)}$$

Now reduces to

$$Sim = \frac{(-1 + 1)}{(1 + 1)} = 0$$

This reduces the similarity measure to a value equal to 0 in the worst case situation.

Average Case:

In average case, $N_{avg} = 0$. The similarity function

$$Sim = \frac{(N_{avg} + \lambda)}{(\lambda + 1)}$$

Now reduces to

$$Sim = \frac{(-\lambda + \lambda)}{(\lambda + 1)} = 0$$

This reduces the similarity measure to a value equal to 0 in the average case situation.

Case-3: $\lambda = 2$

$$Sim = \frac{(N_{avg} + \lambda)}{(\lambda + 1)}$$

$$= \frac{(N_{avg}+2)}{3}$$

In best case, if N_{avg} = 1, then similarity between text files, Sim = 1.

In worst case, if N_{avg} = -1, then similarity between text files, Sim = 0.33

In average case, if N_{avg} = 0, then similarity between text files, Sim = 0.67.

Hence the value of λ is chosen to fit to a value equal to 1

4. CASE STUDY

Consider the document-feature matrix in Table.1. For making the discussion simple, we choose 9 text documents and 10 features obtained after preprocessing phase.

Table.1 : Matrix in Frequency Form

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
File -1	0	1	0	0	1	1	0	0	0	1
File-2	0	0	0	0	0	2	1	1	0	0
File-3	0	0	0	0	0	0	1	0	0	0
File-4	0	0	1	0	2	1	2	1	0	1
File-5	0	0	0	1	0	1	0	0	1	0
File-6	2	1	1	0	0	1	0	0	1	0
File7	3	2	1	3	0	1	0	1	1	0
File-8	1	0	1	1	0	1	0	0	0	0
File-9	1	1	1	1	0	0	0	0	0	0

All these features together represent the global feature set over which the document-feature matrix is formed. This global word set is obtained after initial preprocessing phase. We maintain the matrix in both the frequency form and binary form as shown in Table 1 and Table.2. After applying SVD decomposition, we get the matrices as

Table.2: Matrix in Binary Form

Doc X Doc matrix =

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
File-1	0	1	0	0	1	1	0	0	0	1
File-2	0	0	0	0	0	1	1	1	0	0
File-3	0	0	0	0	0	0	1	0	0	0
File-4	0	0	1	0	1	1	1	1	0	1
File-5	0	0	0	1	0	1	0	0	1	0
File-6	1	1	1	0	0	1	0	0	1	0
File-7	1	1	1	1	0	1	0	1	1	0
File-8	1	0	1	1	0	1	0	0	0	0
File-9	1	1	1	1	0	0	0	0	0	0

```
-0.2466   0.3313  -0.7232   0.2924  -0.1424   0.2045
-0.1949   0.4147   0.4764   0.0311   0.2369   0.1634
-0.0276   0.1813   0.2200  -0.1854   0.1545   0.0741
-0.3494   0.6595  -0.0172  -0.2989  -0.1470  -0.1712
-0.2448  -0.0822   0.2818   0.6803  -0.3797   0.0970
-0.4250  -0.2021  -0.1816   0.1411   0.5520  -0.5973
-0.5492  -0.1943   0.1609   0.0842   0.2575   0.4104
-0.3631  -0.1911   0.1748  -0.2173  -0.5946  -0.4295
-0.3231  -0.3586  -0.1802  -0.5028  -0.0965   0.4214
```

Eigen Value Matrix =

```
4.5519        0        0        0        0        0
     0   2.6313        0        0        0        0
     0        0   1.7572        0        0        0
     0        0        0   1.5635        0        0
     0        0        0        0   1.2577        0
     0        0        0        0        0   0.9461
```

Word x Word matrix =

```
-0.3648  -0.3596  -0.0149  -0.3164   0.0942  -0.2061
-0.3392  -0.1610  -0.5259   0.0095   0.4537   0.4641
-0.4415  -0.1089  -0.0247  -0.5076  -0.0227  -0.3871
-0.3252  -0.3140   0.2488   0.0284  -0.6467   0.5278
-0.1309   0.3765  -0.4214  -0.0042  -0.2301   0.0352
-0.5213   0.2796   0.0978   0.4560  -0.1727  -0.3410
-0.1256   0.4772   0.3865  -0.2899   0.1944   0.0701
-0.2402   0.3344   0.3529  -0.1174   0.2763   0.4255
-0.2678  -0.1819   0.1486   0.5792   0.3418  -0.0950
-0.1309   0.3765  -0.4214  -0.0042  -0.2301   0.0352
```

Consider the absolute values of first column of the word matrix which when sorted gives the words in the order of their importance. This is shown in Table.3 below. The Table.3 below shows both the values of word vectors before and after sorting. After sorting word w_6 is most significant word and word w_7 is the least significant word.

Table.3: Sorting Word Vectors

Before Sorting		After Sorting	
W_1	0.3648	W_6	0.5213
W_2	0.3392	W_3	0.4415
W_3	0.4415	W_1	0.3648
W_4	0.3252	W_2	0.3392
W_5	0.1309	W_4	0.3252
W_6	0.5213	W_9	0.2678
W_7	0.1256	W_8	0.2402
W_8	0.2402	W_5	0.1309
W_9	0.2678	W_{10}	0.1309
W_{10}	0.1309	W_7	0.1256

A. Finding Top-k words

Once we know the significant words from the above step, we can eliminate all the insignificant words from the global feature set which were obtained after the initial preprocessing stage. This may be done by choosing top-k most significant words. We obtain the top-k words by retaining 90% energy from the Eigen values of word vectors. The Table.4 below shows the Eigen values obtained for the matrix of Table.1

Table 4. Eigen Values

	Value
S1	**4.5519**
S2	**2.6313**
S3	**1.7572**
S4	**1.5635**
S5	**1.2577**
S6	**0.9461**
S7	**0.8920**
S8	**0.7342**
S9	**0.1103**

The total energy of the Eigen matrix is summation of all Eigen values = 14.4442. The table.5 shows the energy for top-k Eigen energy values where k is 6, 7, and 8,9,10. To retain 90% of energy we consider top 7 words, hence we reduce the dimensions of the initial matrix of Table.1 from 10 words to 7 words and re-construct the term-frequency matrix as below in Table.6.

Similarly, we may obtain the corresponding binary matrix from the Table.6.

Table 5. Eigen Values

Top-k Eigen Values	Value	% Energy Retained
Top-10	14.4442	100%
Top-9	14.444	100%
Top-8	14.3339	99.24%
Top-7	13.5997	94.15%
Top-6	12.7077	87.98%

Table.6. Reduced matrix in frequency form

	W_6	W_3	W_1	W_2	W_3	W_9	W_8
File1	1	0	0	1	0	0	0
File2	2	0	0	0	0	0	1
File3	0	0	0	0	0	0	0
File4	1	1	0	0	0	0	1
File5	1	0	0	0	1	1	0
File6	1	1	2	1	0	1	0
File7	1	1	3	2	3	1	1
File8	1	1	1	0	1	0	0
File9	0	1	1	1	1	0	0

Table 7. Reduced matrix in binary form

	W_6	W_3	W_1	W_2	W_3	W_9	W_8
File1	1	0	0	1	0	0	0
File2	1	0	0	0	0	0	1
File3	0	0	0	0	0	0	0
File4	1	1	0	0	0	0	1
File5	1	0	0	0	1	1	0
File6	1	1	1	1	0	1	0
File7	1	1	1	1	1	1	1
File8	1	1	1	0	1	0	0
File9	0	1	1	1	1	0	0

This is shown in the Table.7. We compute the standard deviation for the most significant words of the original global feature set instead of all the words in global feature set. This computed standard deviation for each word as shown in Table.8 is later used when computing the similarity degree between any two text files by using the proposed measure. The standard deviation of each word represents the statistical distribution of the corresponding word.

The Table.8 below shows the standard deviation of each feature obtained from Table.7.

Table 8. Standard Deviation matrix

S.Dev	W_6	W_3	W_1	W_2	W_4	W_9	W_8
	0.35	0.53	0.52	0.52	0.52	0.52	0.52

In this section, we consider clustering text documents d_1 through d_9 represented in Table.7. The Table.12 below shows the computation of the similarity matrix which is used to perform text clustering. The table.10 and Table.11 gives the computation values of numerator and denominator of function, N_{AVG}. The similarity between all the combination file pairs from d_1 through d_9 is shown below.

Table 10. Numerator Values of N_{AVG}

α_{12}	0.95
α_{13}	0.97
α_{14}	0.92
α_{15}	0.93
α_{16}	1.92
α_{17}	1.87
α_{18}	0.89
α_{19}	0.92
α_{69}	2.95

Table 11. Denominator Values of N_{AVG}

β_{12}	3
β_{13}	2
β_{14}	4
β_{15}	4
β_{16}	5
β_{17}	7
β_{18}	5
β_{19}	5
β_{69}	6

After applying the clustering algorithm, finally, the clusters formed are
Cluster-1 :< File6, File7>
Cluster-2: <File8, File9>
Cluster-3: <File1, File2, File3, File4>
Cluster-4: < File5>

In the Table.12, the symbol -- indicates those values are not of interest and may be discarded. This is because the similarity measure is symmetric.

5. Text Clustering Procedure

Initially, the clustering process starts by choosing the first maximum value of similarity matrix. The corresponding rows and columns are the members of the clusters formed. The rows or columns indicates index of respective files. The clusters are formed by considering the maximum value at each stage of iteration and discarding those columns which are part of cluster formed newly. The rows are however retained. This is done at each iteration. i.e at each iteration of clustering process, the most maximum value is chosen from the similarity matrix retained at that stage. Finally, when only one file exists or all files have been clustered and there is no other file which may be clustered, the clustering process is terminated and the resulting clusters formed are output.

6. Text Classification

For the purpose of text classification, we choose the similarity measure designed. Consider the document–word matrix n Table.13. If we need to classify any new text document file, we may use the proposed measure to perform text document classification. We may apply the dimensionality reduction to the document-feature matrix and also the test input to reduce the dimensionality and then use these reduced representations to perform classification using the proposed similarity measure. If dimensionality reduction is not required then we may apply the similarity measure directly to perform text classification. To apply the similarity measure the matrix shown in Table.14 must be transformed to binary form. Also, the test document must be transformed to equivalent binary form. For example, consider the following new test document

New document= [6 5 4 3 2 2 0 3 0 1]

To perform test classification, we must transform this text document in to equivalent binary representation

NewDocument = [1 1 1 1 1 1 0 1 0 1]

Table 12. Similarity Matrix

	F2	F3	F4	F5	F6	F7	F8	F9
F1	0.65	0.74	0.62	0.62	0.69	0.63	0.59	0.59
F2	x	0.74	0.66	0.62	0.66	0.63	0.59	0.49
F3	x	x	0.49	0.49	0.49	0.49	0.49	0.49
F4	x	x	x	0.59	0.66	0.71	0.69	0.58
F5	x	x	x	x	0.66	0.71	0.69	0.58
F6	x	x	x	x	x	0.85	0.74	0.75
F7	x	x	x	x	x	x	0.78	0.78
F8	x	x	x	x	x	x	x	0.80
F9	x	x	x	x	x	x	x	x

Table.13 : Matrix in Frequency Form

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
File -1	4	6	2	3	3	1	1	1	0	1
File-2	5	5	3	1	1	2	3	3	2	1
File -3	2	3	4	0	0	0	2	1	2	0
File-4	2	2	3	5	6	4	3	2	1	0
File-5	1	0	1	2	3	2	2	0	2	1
File-6	3	2	0	5	6	5	4	3	0	1
File-7	0	0	2	3	2	3	5	4	6	1
File -8	2	3	3	0	0	3	5	5	4	0
File -9	1	1	0	3	3	2	4	3	3	1

Let the file 1 belongs to class label, easy, files 2,4,6,9 belong to class label medium , remaining files have class label as hard. To perform text classification, we must compute the similarity value for this new test document with all the currently existing text documents.

This is shown in Table.14. Here, f_l to f_9 are features of the global feature set.

In the Table.12, the symbol -- indicates those values are not of interest and may be discarded. This is because the similarity measure is symmetric.

5. Clustering Procedure

Initially, the clustering process starts by choosing the first maximum value of similarity matrix. The corresponding rows and columns are the members of the clusters formed. The rows or columns indicates index of respective files. The clusters are formed by considering the maximum value at each stage of iteration and discarding those columns which are part of cluster formed newly. The rows are however retained. This is done at each iteration. i.e at each iteration of clustering process, the most maximum value is chosen from the similarity matrix retained at that stage. Finally, when only one file exists or all files have been clustered and there is no other file which may be clustered, the clustering process is terminated and the resulting clusters formed are output.

6. Text Classification

For the purpose of text classification, we choose the similarity measure designed. Consider the document–word matrix n Table.13. If we need to classify any new text document file, we may use the proposed measure to perform text document classification. We may apply the dimensionality reduction to the document-feature matrix and also the test input to reduce the dimensionality and then use these reduced representations to perform classification using the proposed similarity measure. If dimensionality reduction is not required then we may apply the similarity measure directly to perform text classification. To apply the similarity measure the matrix shown in Table.14 must be transformed to binary form. Also, the test document must be transformed to equivalent binary form. For example, consider the following new test document

New document= [6 5 4 3 2 2 0 3 0 1]

To perform test classification, we must transform this text document in to equivalent binary representation

NewDocument = [1 1 1 1 1 1 0 1 0 1]

Table 12. Similarity Matrix

	F2	F3	F4	F5	F6	F7	F8	F9
F1	0.65	0.74	0.62	0.62	0.69	0.63	0.59	0.59
F2	x	0.74	0.66	0.62	0.66	0.63	0.59	0.49
F3	x	x	0.49	0.49	0.49	0.49	0.49	0.49
F4	x	x	x	0.59	0.66	0.71	0.69	0.58
F5	x	x	x	x	0.66	0.71	0.69	0.58
F6	x	x	x	x	x	0.85	0.74	0.75
F7	x	x	x	x	x	x	0.78	0.78
F8	x	x	x	x	x	x	x	0.80
F9	x	x	x	x	x	x	x	x

Table.13 : Matrix in Frequency Form

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
File -1	4	6	2	3	3	1	1	1	0	1
File-2	5	5	3	1	1	2	3	3	2	1
File -3	2	3	4	0	0	0	2	1	2	0
File-4	2	2	3	5	6	4	3	2	1	0
File-5	1	0	1	2	3	2	2	0	2	1
File-6	3	2	0	5	6	5	4	3	0	1
File-7	0	0	2	3	2	3	5	4	6	1
File -8	2	3	3	0	0	3	5	5	4	0
File -9	1	1	0	3	3	2	4	3	3	1

Let the file 1 belongs to class label, easy, files 2,4,6,9 belong to class label medium , remaining files have class label as hard. To perform text classification, we must compute the similarity value for this new test document with all the currently existing text documents. This is shown in Table.14. Here, f_1 to f_9 are features of the global feature set.

Table. 14 New Document Similarities

Existing File	Similarity with new-file
File-1	0.944
File-2	0.850
File-3	0.697
File-4	0.849
File-5	0.799
File-6	0.888
File-7	0.799
File-8	0.748
File-9	0.849

The new test document file has maximum similarity value w.r.t file-1 as shown in Table.15. So it is classified to the corresponding class of first file as easy.

7. Conclusions

In this work, we use the concept singular valued decomposition to perform dimensionality reduction and use this reduced dimensionality text documents to perform text classification and text clustering. To perform text clustering, we make use of the proposed distance measure which is the improved version of our previous measure which does not consider the distribution of features of the document. The clustering approach is performed to cluster the text documents with the proposed similarity measure. For text classification, we use the proposed similarity measure and classify the new text document to the corresponding class label of training dataset. The proposed measure is designed by considering, worst case, average and best case situations and validated formally.

References

[1] Andrew Stranieri, John Zeleznikow. Information retrieval and Text Mining. Knowledge Discovery from Legal Databases Law and Philosophy Library Volume 69(2005), 147-169

[2] Hussein Hashimi, Alaaeldin Hafez, Hassan Mathkour, Selection criteria for text mining approaches, Computers in Human Behavior, Volume 51, Part B, October 2015, Pages 729-733, ISSN 0747-5632

[3] Yannis Haralambous and Philippe Lenca: Text Classification Using Association Rules, Dependency Pruning and Hyperonymization. Proceedings of DMNLP, Workshop at ECML/PKDD, Nancy, France, 2014.

[4] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo: Text mining for market prediction: A systematic review. Expert Systems with Applications 41 (2014) 7653–7670.

[5] Sajid Mahmood, Muhammad Shahbaz, and Aziz Guergachi, "Negative and Positive Association Rules Mining from Text Using Frequent and Infrequent Itemsets," The Scientific World Journal, vol. 2014, Article ID 973750, 11 pages, 2014. doi:10.1155/2014/973750

[6] Wen Zhang, Taketoshi Yoshida, Xijin Tang, Qing Wang: Text clustering using frequent item sets. Knowledge-Based Systems 23 (2010) 379–388.

[7] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu: Effective Pattern Discovery for Text Mining. IEEE Transactions on Knowledge and Data Engineering. Volume 24, No. 1, Jan 2012

[8] Fodor, I.K. (2002) A Survey of Dimension Reduction Techniques. Technical Report, UCRL-ID-148494, Lawrence Livermore NationalLaboratory, Livermore. http://dx.doi.org/10.2172/15002 155.

[9] Christopher J. C. Burges: Dimension Reduction: A Guided Tour. Foundations and Trends in Machine Learning Vol. 2, No. 4 (2009) 275–365.

[10] Yung-Shen Lin; Jung-Yi Jiang; Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol.26, no.7, pp.1575-1590, July 2014.

[11] Sunghae Jun et.al. Document clustering method using dimension reduction and support vector clustering to overcome sparseness, Expert Systems and Applications, 41(2014),3204-3212.

[12] Jung-Yi Jiang et.al A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification, IEEE Transactions on Know-ledge and Data Engineering,Vol.23,No.3, 2011

[13] Shuqing Huang. 2007. A Comparative Study of Clustering and Classification Algorithms. Ph.D. Dissertation. Tulane University, New Orleans, LA, USA. Advisor(s) Parviz Rastgoufard. AAI3258261.

[14] Hui Han, Eren Manavoglu, C. Lee Giles, Hongyuan Zha: Rule-based word clustering for text classification. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval.

[15] Lam Hong Lee, Chin Heng Wan, Rajprasad Rajkumar, and Dino Isa. 2012. An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization. Applied Intelligence 37, 1 (July 2012), 80-99.

[16] Libiao Zhang, Yuefeng Li, Chao Sun, and Wanvimol Nadee. 2013. Rough Set Based Approach to Text Classification. In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 03(WI-IAT '13), Vol. 3. IEEE Computer Society, Washington, DC, USA, 245-252.

[17] Chin Heng Wan, Lam Hong Lee, Rajprasad Rajkumar, and Dino Isa. 2012. A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. Expert Syst. Appl. 39, 15 (November 2012), 11880-11888.

[18] Sofus A. Macskassy and Haym Hirsh. 2003. Adding numbers to text classification. In Proceedings of the twelfth international conference on Information and knowledge management (CIKM '03). ACM, New York, NY, USA, 240-246.

[19] Dell Zhang and Wee Sun Lee. 2006. Extracting key-substring-group features for text classification. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '06). ACM, New York, NY, USA, 474-483.

[20] Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2008. Text classification based on multi-word with support vector machine. Know.-Based Syst. 21, 8 (2008), 879-886.

[21] Shuigeng Zhou and Jihong Guan. 2002. An Approach to Improve Text Classification Efficiency. In Proceedings of the 6th East European Conference on Advances in Databases and Information Systems (ADBIS '02), 65-79.

[22] Berna Altınel, Murat Can Ganiz, and Banu Diri. 2015. A corpus-based semantic kernel for text classification by using meaning values of terms. Eng. Appl. Artif. Intell. 43, C (August 2015), 54-66.

[23] Jing Gao and Jun Zhang. 2005. Clustered SVD strategies in latent semantic indexing. Inf. Process. Manage. 41, 5(2005), 1051-1063.

[24] Reddy, G.S.; Rajinikanth, T.V.; Rao, A.A., "A frequent term based text clustering approach using novel similarity measure," in Advance Computing Conference (IACC), 2014 IEEE International , vol., no., pp.495-499, 21-22 Feb. 2014

[25] G. SureshReddy, T. V. Rajinikanth, and A. Ananda Rao. 2014. Design and analysis of novel similarity measure for clustering and classification of high dimensional text documents. In Proceedings of the 15th International Conference on Computer Systems and Technologies (CompSysTech '14), 194-201.

[26] Shehata, S.; Karray, F.; Kamel, M.S., "An Efficient Concept-Based Mining Model for Enhancing Text Clustering," in *Knowledge and Data Engineering, IEEE Transactions on* , vol.22, no.10, pp.1360-1371, Oct. 2010.

[27] Yanjun Li; Congnan Luo; Chung, S.M., "Text Clustering with Feature Selection by Using Statistical Data," in *Knowledge and Data Engineering, IEEE Transactions on* , vol.20, no.5, pp.641-652, May 2008.

[28] Papapetrou, Odysseas; Siberski, Wolf; Fuhr, N., "Decentralized Probabilistic Text Clustering," in *Knowledge and Data Engineering, IEEE Transactions on* , vol.24, no.10, pp.1848-1861, Oct. 2012 doi: 10.1109/TKDE.2011.120

[29] Skabar, A.; Abdalgader, K., "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm," in *Knowledge and Data Engineering, IEEE Transactions on* , vol.25, no.1, pp.62-75, Jan. 2013. Doi: 10.1109/TKDE.2011.205

[30] Shie-Jue Lee; Jung-Yi Jiang, "Multilabel Text Categorization Based on Fuzzy Relevance Clustering," in *Fuzzy Systems, IEEE Transactions on* , vol.22, no.6, pp.1457-1471, Dec. 2014 doi: 10.1109/TFUZZ.2013.2294355

[31] G.Suresh Reddy, A.Ananda Rao, T.V.Rajinikanth. ”An improved Similarity Measure for Text Clustering and Classification. Advanced Science Letters, 2015. In press

QoS Web Service Security Dynamic Intruder Detection System for HTTP SSL services

M.Swami Das[1], A.Govardhan[2], D.Vijaya lakshmi [3]

Assoc. Professor, CSE, MREC[1], Professor, SIT,JNTU Hyderabad[2], Professor, Dept. of CSE, MGIT Hyderabad, India[3]

Abstract: Web services are expected to play significant role for message communications over internet applications. Most of the future work is web security. Online shopping and web services are increasing at rapid rate. In this paper we presented the fundamental concepts related to Network security, web security threats. QoS web service security intrusion detection is important concern in network communications and firewalls security; we discussed various issues and challenges related to web security. The fundamental concepts network security XML firewall, XML networks. We proposed a novel Dynamic Intruder Detection System (DIDA) is safe guard against SSL secured transactions over message communications in intermediate routers that enable services to sender and receiver use Secured Session Layer protocol messages. This can be into three stages 1) Sensor 2) Analyzer and 3)User Interface..

Keywords: Web Security, QoS web service, HTTP, Intruder detection, Secure Socket Layer, Network Security

1.Introduction

Intrusion detection system is a device or software application that monitors malicious attacks or network traffic if any policy violations [1]. Web services applications communicate and coordinate message passing between client and server. A web service provides functionality and services to the web users. The users to communicate in network channels, the hacker or Intruder tries to operate various attacks such as, DDOS attack, side channel attack, authentication attack, man in the middle attack, cloud computing attacks and steal sensitive information. Hacker execute arbitrary or malicious code in the system due to vulnerability, weak security and no Intruder Detection and Monitoring system.[2]. In Intrusion detection system is a device or software application that monitors malicious attacks or network traffic if any policy violations [1]. A web service provides functionality and services to the web users. The users to communicate in network channels, the hacker or Intruder tries to operate various attacks such as, DDOS attack, side channel attack, authentication attack, man in the middle attack, cloud computing attacks and steal sensitive information. Hacker execute arbitrary or malicious code in the system due to vulnerability, weak security and no Intruder Detection and Monitoring system.[2]. In recently ISRO website homepage hacked by hackers, other examples related to Government and other web sites discussed[3,4]. It is essential to provide Intruder Detection and monitoring system for Government Institutions, Diplomatic offices, Energy, oil and gas companies, Research Institutions, private equity firms, and activist. Frequently to monitor and control the valid and authorized data operations over the network.

Web security has three important concepts confidentiality, integrity and availability. Confidentiality means Information not available to unauthorized users. Integrity defined by the property that data has not been modified by unauthorized users, and availability means web services are accessible to authorized users with access restrictions.[5] Intrusion: attempting to attack into or misuse the system from outside network or legitimate users of the network, intrusion can be a physical, system or remote intrusion. Automatic Intrusion detection system sensor, Analyzer and user interface. Intrusion Detection systems can be classified as i)Anomaly detection ii) signature based misuse iii) host based iv) network based v) stack based
The rest of this paper is organized as follows Section 2: Related work, Section: 3 Issues and challenges, Section 4: Web security and Network

security, Section: 5 Discussions and Interpretations and Section 6. Conclusion.

2. Related Work

Zhiwen Bai etl, Proposed DTAD, a dynamic taint analysis detector aiming to protect malicious attacks and vulnerabilities. Attacker process is detected and precision intrusion, signature of collection of virtual systems and comparing network data and log files used to identify the attacks.[6]

Jiang Du etl, studied man in the middle attacker use ARP deception for both sides communication. Man in the middle will generate own public, private and self digital certificate, and this is interactive process validated by Service provider.[7]

Taro Ishitaki,etl proposed intrusion detection system using Neural network, Fuzzy logic, Probabilistic reasoning, Genetic algorithms capable for finding pattern behavior to detect normal and attack conditions[8].

The SOAP messages to ensure integrity and authentication during the data transmission. Web services require partial signing of SOAP request which is achieved using XML signature by WSDL documents and operations as suggested by Padmanabhuni and Adarkar etl[9,10]. Web service security is critical task for message invocations by web servers. SOAP uses XML encryption, XML digital signature, SSL/TLS methods. XML message security is achieved by service oriented security functionality.

Web service standards SOAP level security authentication, authorization management. Web security is defined as attach signature and encryption header to SOAP messages. It describes security tokens. Web security policy is defined as set of specifications that describe rules, constraints and other business policies on intermediaries and end points. (Example. Encryption algorithms).Web security trust describes a frame work to design a model that enables web services to securely inter-operate request, issues, and exchange security operations.

2.1 XML Firewall

Web services environment, malicious attacks and DoS attacks are new challenges. Firewall allows to the Service providers residing in a network to be invoked from outside the network, and keeping a high security[9]. HTTP protocol is not suitable for creating public key infrastructure. The prototype is used by application behind a firewall.[11]

2.2 XML networks

web service management vendors develop network based solution for web service applications to provide better QoS web services with security to various networks endpoints service consumers and service providers.[11].

Fang Qi .etl, .proposed Automatic Detecting Security Indicator (ADSI) for preventing Web spoofing on a confidential computer which is a harmless environment. It creates a random indicator to identify and detect bogus pages with URL screening data.[13]

Jaing Du,.etl analyzed as a case study secured socket layer man in the middle attack based on SSL certification interaction. Attacker place computer gives a vital role two communication processes. [14]

Lin-Shung Huang., etc introduced a new method for detecting SSL man -in- the middle attacks against website users, over of SSL connections at the top web sites by checking certificates as number of CA certificates. Trace any malware in SSL connections for identify and provide better protection. [15]

3 Issues and challenges

The following are the list of issues /challenges in Web security/network security. Digital certificates are designed to establish credentials of the people use Router configurations with weak vulnerabilities and security policies described in Table.1. Web security developers provide secured operations and safety steps necessary to identify trusted systems. [16]

Table.1. Router or firewall configurations with weak or vulnerabilities

Web services Solutions or threats	Problem in domain or Safety precautions
Web service has arbitrary disclosure policy	Provide strong policies to web services
Passwords stored in browser	Do not save passwords in browser history
Institutions, organizations malicious code attacks, virus	Web security , Frequently monitor network operations. Use SSL security

Malware, Denial of service attacks to modems / routers against other systems by unknown users by stealing personal information and credentials to access certain web sites.

Hackers used stolen laptops/equipment to hack web data where there is vulnerability like private wireless network or wireless network is unsecured with no password is immediately accessible to hackers. Hacker used wireless antenna and software nearby buildings and capture/ steal information like passwords, email-messages, and any data transmitted over the network when a network is not secured. [5].Hacker will use some tools described in table.2. Brute-force attack is the password cracking method, trying all the solutions seeking one fits[11].Stealing the login password controlling the devices by malicious scripts and malicious DNS servers attacked on DSL modems.

3.1 Man-in-the-Middle (MIM) attacks

This attack where the attacker secretly relays and possibly alters the communication between two parties who believe they are openly communicating with each other. Attacker intercept all relevant messages by passing between victims and adding extra information. Attackers trying to access the services using fake address, fake certifications. Examples of MIM attacks One provides free Wi-Fi service with malicious software.

3.1.1 ARP Cache Poisoning

Sender and receiver over message communication, PC sends IP packets broad cast to all systems in subnet. ARP(address resolution protocol is not secured protocol).

3.1.2 DNS Spoofing

DNS cache poisoning is a computer hacking attack, where by data is communicated into a Domain Name System (DNS) resolver’s cache, causing the name server to return an incorrect IP address, diverting traffic to the attacker’s computer (or any other computer).Attackers creating a fake web site by redirecting data to shadow servers.

3.1.3 Session Hijacking: Client to server when session established, the hacker capture cookies information and diverting the session communications to un-trusted systems

3.1.4 Session hijacking attack

Communication over TCP connections. Session normally consists of string of variables used in URL stealing and predicting valid session token to gain unauthorized access to the web server [17]

Table.2.Tools and software’s used to steal the data

Web services Solutions or threats	Problem in domain or Safety precautions
Suspicious downloads or plugins	Use firewall in secure network
Terminals with chip card vulnerabilities	Alert any where service by authentication and secret key.

4.Web Security Network security

Web Service Security: Three types of digital certificates are domain validated certificate, organizational validation certificate and Extended validation certification. Domain validated Certificate: trusted domain name of owner. Organizational Validation Certificate: validation of organization by DNS names. Extended validation certification: Certificate Agent must meet minimum validation criteria. Organizations, application vendors, Browser makers issue extended validation certificate.[4] Web services standards worked at w3C, OASIS, IETF and other bodies to enable faster inventions of web services and security. A web service provides a flexible set of mechanism to design a range of security protocols. It is essential to design non-vulnerable protocols for web services security. Web services specifications goals to provide multiple security token formats, multiple trust domains, multiple signature formats, multiple encryption methodologies, and end to end message content security.[12]

4.1 Intruder Identification and Detection System

4.1.1 Various Attacks

Unauthorized system used to attack on router or servers using various attacks (DDOS attack, side channel attack, Man in the middle attack , Authentication attack and cloud computing attacks) methods practiced due to various reasons like, not secured web site, malicious code, denying encrypt , weak secret keys, vulnerabilities in content security, and policy constraints. In Figure.1. shows the intruder attacks on router.

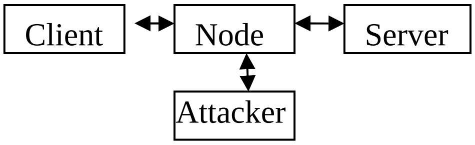


Figure.1. Intruder attacks on services

4.1.1.1 Denial of Service attacks

DDOS attack were launched from distributed attacking hosts. This is launched two phases. First an attacker builds a network which is distributed and consists of thousand of compromised computers are called(Zombies, attacking hosts). The attacker hosts flood of tremendous volume of traffic towards victims either under command or automatically [32].

4.1.1.2 Attack to change DNS settings

Attackers directly targeting DNS server two ways Cybersquatting aim is to steal the Victims identity and or divert traffic from victims website. Name jacking or theft : by appropriate the domain name (updating the holders field or taking control) by technical means to divert the traffic such as modifying the name of hosting the site.[18]

4.1.1.3 Authentication attack

This type of attack targets and attempt to take advantage of following Brute force : allow attacker to guess persons username, other credentials by using Automated trail and error Insufficient Authentication: Allows an attached to access a web site sensible information without having to properly authenticate in web site. Sending phishing mail to user to steal sensitive information[19]

4.2 Intruder Detection System

Intruder Detection System has two type namely Network Intrusion Detection System and Host based Intrusion Detection System

4.2.1 Network based Intrusion Detection System

It deals with traffic accounting and network flow information. This system is implementing in Routers and switches Input and Output HTTP / TCP data, and testing various functions like port scanning , Reassembling, decoding, detecting virus, protocol violations.

4.2.2 Host based Intrusion Detection System

It deals with Analyzing logging facility for almost all failed or success services. The system is implementing in Routers or Firewall to access authorized client. It calculates the cryptographic checks of files, including owner ,group changes, and also checks system integrity.[20]

Web services accessed by sending SOAP messages to endpoints. This is handed by transport layer security protocol such as HTTP, SSL, and TLS others. This ensures secured peer to peer messages. Web based security standards mapping to XML message security. All protocols use to carry security data as part of XML document. The XML document is critical part of security requirement of web services. [9]

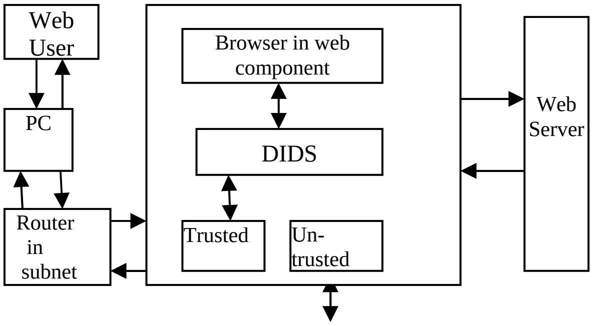


Figure.2. Dynamic Intrusion Detection System

4.3 Secure Socket Layer and Transport Layer

Security HTTP Secured Socket Layer Protocol: HTTP over Secured Socket Layer combination to secure communication between browser and web server systems. SSL Secured Socket Layer Protocol is transient, peer to peer communication, SSL protocol stack link associated with SSL session Record Protocol operation. These SSL sessions in association between client & server by handshake protocol, with defined set of cryptographic parameters that may be shared by multiple SSL connections. [3, 12, 13]. HTTP protocol stack provides transfer information for web services interaction can operate on top of SSL. Three layers are defined as part of SSL such as Hand shake protocol, The change of cipher spec protocol and the Alert protocol. These protocols are used in management of SSL exchange.

4.4 Proposed Model: Dynamic Intruder Detection system

The Automated Intruder Detection System shown in Figure. 2. It will detect the unauthorized or hacker requests by invoking a procedure Intrusion Detection System in four subpaths, that are user requests to subnet router point to point in Transport layer, browser in the Intrusion detection system detection system invokes a procedure to check, Certification, digital signature of trusted client. If trusted request as a result then it inserts the process for further processing into Deque. The deque holds a batch of trusted services routed to next hop via point to point protocol.In subpath3 browser contents security not known to attacker by pedlock security. The forth sub-path content in the web server connecting a session request for web services. The algorithm:1,2 and 3 depicted table. Intrusion Detection System Message Format alert:(messageid; create time;nt pstamp;date;time; source;node;address;message; flag)

4.4.1 Components of DIDS

Dynamic Intrusion Detection System has three components are sensor, Analyzer and User interface. Overall network security maintains a security state. when threat occurs by executing an event, the system will check the context of
the event and data by following

4.4.1.1 Sensor

Sensor are responsible for collecting data. Example network packet, log files, and system call traces, sensor collect and forward to the analyzer

4.4.1.2 Analyzer

Analyzer receiver input from one or more sensors from the system . Control the behavior of the system.

4.4.1.3 User Interface

The user interface to DIDS that enables a user to view output from the system or control the behavior of the system. System component as manager or console component.[20]

```
Algorithm: 1 Initialization of Request
Procedure: DYNAMIC INTRUDER
DETECTON  SYSTEM
Input : Sensor/ node send a service request
Output: Trusted service or Un-trusted service
begin
1. Establish connection between sender and receiver
2. User system to web server consists of four sub paths
3. subpath1: user request to router in subnet(trusted system)
4. Subpath2:Web browser in router checks the procedure using Dynamic Intruder detection system.
5. Identify the request process trusted request pushed onto Deque and un-trusted requests rejected and access restricted.
6. Subpath3: Web browser content and security sign which are not known to the attacker. Ex icon with Padlock security sign unknown to the attacker.
7. Subpath4:Web content to Web server: Connecting via subnet routers with trusted systems the request connection established between sender to receiver
      using SSL handshake protocol
   end
```

```
Algorithm.2: Connection Establishment.
Procedure PROCESS DETECTING TRUSTED REQUEST
Input :Web service request
Output : Secured HTTP Session layer
Begin
 1. Read  DIDS
 2. Client sends a request to Web server by invoking HandShake protocol using cryptographic parameters (clientid, clientMAC, ClientSecretKey, serverid)
 3. During handshaking protocol session is created successfully by resuming the  new state if already the state is running. With Session identified by its state, prior to encryption algorithms.
 4. Each connection creates a secure session layer and sets the flag. Here flag indicated the connection.
 5. The request process is checking by Certificate Agent, and Digital certificate.
 6.  Detecting trusted service or un-trusted service. if request is trusted service then insert the process in Deque for further processing communication to next  hop if connection request is un-trusted requests are denied/ rejected
    end
```

```
Algorithm.3 Closing the Connection
Procedure WS SECURED CONNCTION
        Input: web service request message
   Output: Secure access control by encryption
Begin
1. Read PROCESS DETECTING TRUSTED REQUEST
2. User data is verified with the data with existing data of concerned  web server.
3. If(HTTPrequest is successful) then Connection is established, under service access policy else Connection is closed with notification
4. if connection is established enable decryption of data at the web server
5. Message communication is accomplished by SSL encryption method.
6.  close the connection
7. Connection closure if connection is closed in HTTP record
8. TLS level exchange close notify alert then close TCP connection
9. handle TCP close before alert exchange send or completed
End
```

5 Discussions and Interpretations

Secure Socket Layer provides security services between TCP and applications that use TCP. Internet standards TLS, SSL/TLS provide confidentiality using symmetric encryption and message integrity and Authentication code. This DIDS (Dynamic Intruder Detection System) protecting the man-in the middle attacks and deny services. And allows trusted services forwarded to next hop to reach peer entity web server and Web service applications. The discussions and interpretations for web securities, precautions and remedies are described in table.2 provides the information related to web security/network security problems and proposed solutions/precautions to meet network securities QoS parameters such as confidentiality, integrity, data authentication, and availability of information to trusted users from web service systems to detect various attacks. The genuine merchants by Digital certificates and required policy constraints to validate authentication process in DIDS system architecture.

6 Conclusions

Web services are expected to play increasing important role for message communications over internet applications. Most of future work is web security. Online shopping and web services are increasing in the world. In this paper we described the fundamental concepts related to web security threats, web server architectures, web server protocols. QoS web service security is important concern in network communications. Firewalls security, various issues and challenges of web security. Discussed fundamental concepts, network security encryption and decryption process, and Network security hierarchies.

We proposed a novel Dynamic Intruder Detection System(AIDA) is safe guard against SSL secured transactions over message communications to intermediate routers that enable services to sender and receiver use Secured Session Layer protocol messages. This can be into three stages 1) Weak security assumption 2) Intruder attacks on browser 3)Trusted system detect service and safe guard information. As a case study we proposed the architecture of system in Figure .2.In future we can extend this paper to E-Commerce, Online Financial transactions, and this security concepts used for designing and developing Firewalls which will protect web services applications.

References

[1] https://en.wikipedia.org/
[2] www.livehacking.com
[3] http://www.thehindu.com/news/national
[4] http://www.ndtv.com/topic/websites-hacked
[5] M.Swami Das, A.Govardhan, and D.Vijya lakshmi: QoS web service Security Access Control case study using HTTP Secured Socket Layer Approach ICEMIS 15, September 24-26, 2015, Istanbul, Turkey 2015 ACM.
ISBN 978-1-4503-3418-1/15/09
[6] Zhiwen Bai,Liming Wang, Jinglin Chen,Jain Liu,Xiyang Liu on ” DTAD A Dynamic Taint Analysis Detector for Information Security”,IEEE, Web age Information system 2008, pp,591-597
[7] Jaing Du,Xing Li and Hua Huang :A study of man in the middle attack based on SSL certificate interaction”,IEEE,ICIMCCC 2011, pp 445-448
[8] Taro Ishitaki, Donald , Yi Liu, Tetsuya Oda, Leonard Barolli, and Kazunori Uchida : Application of Neural Networks for Intrusion Detection in Tor Networks.IEEE, ICAINAW 2015, pp 67-72
[9] https://events.ccc.de/congress/2005/fahrplan /638-22c3 ids.pdf
[10] www.cs.ucsb.edu
[11]Service-oriented Software System Engineering: Challenges
and Practices by Z Stojanovi, Ajantha D
[12]Service-oriented Software System Engineering: Challenges and Practices by Z Stojanovi, Ajantha D
[13] Fang Qi, Zhe Tang, Guojun Wang on” Attacks vs. Countermeasures of SSL Protected Trust Model”, IEEE confernece 2008, pp1886-1991
[14] Jin-Ha Kim, Gyu Sang Choi and Chita R. Das :A Load Balancing Scheme for Cluster-based Secure Network Servers,IEEE
[15] Lin-Shung Huang, Alex Ricey, Erling Ellingseny,Collin Jackson :Analyzing Forged SSL Certificates in the Wild,2014 IEEE Symposium on Security and Privacy,pp.83-97
[16] Neal Leavitt :Internet Security under Attack: The Undermining of Digital Certificates”,Technology news in IEEE 2011,pp17-20
[17] https://en.wikipedia.org
[18] https://www.afnic.fr/ DNS Types of attack and security techniques
[19]http://www1.ibm.com/support/knowledgecenter/SSB2MG.6.0/com.ibm.ips.doc/concepts/wap authentication.htm
[20] https://s2.ist.psu.edu/paper/ddos-chap-gu-june-07.pdf

Does Software Structures Quality Improve over Software Evolution? Evidences from Open-Source Projects

Mamdouh Alenezi and Mohammad Zarour
College of Computer & Information Sciences
Prince Sultan University, Riyadh 11586
Saudi Arabia

Abstract

Throughout the software evolution, several maintenance actions such as adding new features, fixing problems, improving the design might negatively or positively affect the software design quality. Quality degradation, if not handled in the right time, can accumulate and cause serious problems for future maintenance effort. Several researchers considered modularity as one of the success factors of Open Source Software (OSS) Projects. The modularity of these systems is influenced by some software metrics such as size, complexity, cohesion, and coupling. In this work, we study the modularity evolution of four open-source systems by answering two main research questions namely: what measures can be used to measure the modularity level of software and secondly, did the modularity level for the selected open source software improves over time. By investigating the modularity measures, we have identified the main measures that can be used to measure software modularity. Based on our analysis, the modularity of these two systems is not improving over time. However, the defect density is improving overtime.

1 Introduction

Software evolve for many reasons that include continuing change, increasing complexity, continuing growth and etc. This means that software need to fix problems, to accommodate new features, and to improve their quality. All these maintenance activities lie within corrective, preventive, adaptive and perfective maintenance that lead to software evolution. In order for the software to survive for a long period, it needs to evolve. This paper is an extended version of our previous work [1]. In this paper we study the software structures quality and investigate more their improvement opportunities over the evolution of four different open source projects.

Software end-users are usually concerned about the external software quality factors depicted as efficiency, usability, and reliability while developers and software engineers are also concerned with the internal quality factors such as evolution and reusability [2]. Software keeps evolving after it has been set in use for the first time. The cost associated with software maintenance and evolution is estimated to be 60% to 80% of total costs associated with a software system [3]. Software evolution is correlated with software structures and complexity [4]; Software structures can be altered via maintenance activities which usually introduce new source code changes that may introduces new dependencies among software elements e.g. packages, methods and classes. Most of the software evolution studies highlight the changes in statistical techniques by analyzing its evolution measures [5], little effort has been carried to comprehend how exactly the structure of these systems evolve [6]. For that reason, we focus in this paper on studying software structures’ quality and investigate their improvements during software evolution. Our investigation is based on open source software systems by considering various object oriented structural software measures.

Software structures refer to the various program elements (modules)that make up certain software. the way these elements are organized in the program defines its structural complexity [7]. Modularity has great effect on software development and evolution [8][9][10]. It plays a central role in the design and production of software artifacts, mainly when developing large and complex software [11]. Modularity is one of the maintainability characteristics of the ISO/IEC SQuaRe quality standard series [12]. According to this standard, modularity is defined as a degree to which a system or computer program is composed of discrete components such that a change to one component has minimal impact on other components [12]. modularization is the process of decomposing a system into logically cohesive and loosely-coupled modules that hide their implementation from each other and offer functionalities to the outside world through a well-defined interface [13, 14]. Maintenance activities during software evolution might negatively or positively affect software quality including modularity, enhancing software modularity will improve the flexibility and understandability of software systems. As software Modularity increase, its complexity decrease. High modularity in open source allows multiple developers to work on the same software entity, usually in competition, which increases the probability of timely, high-quality solutions [15].

Modularity is an essential property of quality software. High modularity improves the flexibility and understandability of the software system [8], whereas low modularity causes costly refactorings and software bugs [10]. Therefore, modularity is usually utilized as an essential criterion for evaluating the software design quality [12]. In this paper modularity measures are used as means to study the software structures quality and their evolution over projects’ releases.

The remainder of this paper is organized as follows: Section 2 discusses the research methodology adopted in this paper. Section 3 states the measures used in this study. The data collection mechanism is given in Section 4. Data analysis and results are presented in Section 5. Threat to validity are discussed in Section 6. Section 7 discusses related work. Conclusions are presented in Section 8.

2 Research Methodology

In this research work we are applying various modularity measures to empirical data taken from open source software. The data are collected from PROMISE, the software engineering repository. Nowadays, open source software repositories provide researchers with the possibility to access large amount of publicly available data for analysis to produce new studies and results. In our study, we will investigate the relationship among various design measures and software modularity. Modularity forms our dependent variable to be studied while the various design measures form the independent variables. Our empirical study focuses on the following research questions:

1.What measure(s) can be used to measure the modularity of OO software programs?

2. Did the modularity of the OO programs studied in this research work improve over the various versions?

To answer these questions, we will follow the following steps:

1. Identify the applicable set of measures related to the modularity (Section 4)
2. Identify the set of open source software systems to be used in this research work and collect necessary data pieces from PROMISE repository needed to calculate the specified measures (Section 5)
3. Analyze and report findings (Section 6)

3 Measures in This Study

Various measures are used to measure the quality of modularization. Although deciding which measures can be adopted in experiments on object oriented software modularity is a hard task [16], we decided to consider coupling, cohesion and complexity as measures to be considered in this study. According to [16] measures related to these three internal attributes are among the most adopted measures by experts in the domain. Coupling is the degree of interdependence between modules, whereas cohesion is the intra-modular functional relatedness which describes how tightly bound the internal elements of a module are to one another [14]. An excessive coupling between a system modules affects its modularity but promoting encapsulation and reducing coupling improve modularity [17]. Complexity is also revealed by both cohesion and coupling. Higher cohesion indicates lower complexity, when coupling increases, the complexity also increases. Coupling, cohesion, and complexity relate strongly to the maintenance effort [18]. Moreover, Defect Density is used as a measure of software product quality to investigate if the defect level is improving over successive releases.

This section presents the definition of the measures used in the study. For more detailed definition about these measures refer to [19]. Modularity measures assess the degree to which a system or computer program is composed of discrete components such that a change to one component has minimal impact on other components. SQuaRE standard defined two basic modularity measures:

- Coupling of components: How strong is the coupling between the components in a system or computer programs? basically two measures are used for coupling measure: Coupling between object classes (CBO) and Response for a Class (RFC).
- Cyclomatic complexity: How many software modules have the acceptable cyclomatic complexity? The cyclomatic complexity is measured by two main measures namely: Weighted Methods per Class (WMC), and McCabe' s Cyclomatic Complexity (CC).

We observe the modularity of open source software systems by measuring coupling, cohesion, and complexity measures. While major emphasis has been on object oriented measures proposed by Chidamber and Kemerer [20], we have also considered other relevant measures related to coupling and cohesions as shown in the following sub-sections.

3.1 Coupling

Beside the two basic coupling measures given above, we have also chosen other measures that measure the interconnection of software modules. this includes: Afferent couplings

(Ca), Efferent couplings (Ce), Coupling Between Methods (CBM). These coupling measures are well-known and were excessively studied in the literature. Accordingly the selected coupling measures include:

- Coupling between object classes (CBO): It represents the number of other classes that are coupled to the current class. This coupling can occur through method calls, field accesses, inheritance, arguments, return types, and exceptions.
- Response for a Class (RFC): RFC is the measure of number of methods that can be invoked in response to a message received by an object of the class. Ideally, RFC should measure the transitive closure of the call graph for each method.
- Afferent couplings (Ca): It represents the number of classes from other packages depending on classes in this package. This describes the packages responsibility. Ca is the number of other packages depending on one package. A high number indicates bad design, or that the package is used for crosscutting concerns.
- Efferent Couplings (Ce): It represents the number of packages the classes of this package depend upon. This describes the packages independence. This can be used to point out non-adherence to the design if certain packages have an unreasonable high number of efferent couplings.
- Coupling Between Methods (CBM): It represents the total number of new/redefined methods to which all the inherited methods are coupled. An inherited method is coupled to a new/redefined method if it is functionally dependent on a new/redefined method in the class. Therefore, the number of new/redefined methods to which an inherited method is coupled can be measured.

3.2 Cohesion

To study software systems cohesion, we have chosen different measures that measure the cohesion of software modules. These cohesion measures are well-known and were excessively studied in the literature. We selected the following cohesion measures:

- Lack of cohesion in methods (LCOM): It counts the sets of methods in a class that are not related through the sharing of some of the class fields. It is calculated by subtracting from the number of method pairs that do not share a field access the number of method pairs that do.
- Lack of cohesion in methods (LCOM3): It is an improved variation of the LCOM measure. It calculates the cohesion of the class by considering the effective usage of the class attributes.
- Cohesion Among Methods of Class (CAM): It computes the relatedness among methods of a class based upon the parameter list of the methods. It sums the number of different types of method parameters in every method and divides it by a multiplication of number of different method parameter types in whole class and number of methods.

3.3 Complexity

To study software system’ s cohesion, we used different complexity measures which are well-known and excessively studied in the literature. These measures include:

- Weighted Methods per Class (WMC): It is the sum of the complexities of all class methods.
- McCabe' s Cyclomatic Complexity (CC): It is equal to the number of different paths (decision points) in a method plus one. We report Avg(CC) which is the arithmetic mean of the CC value in the investigated class.

3.4 Defect DensityEvolution

Defect Density is post-release defects per thousand lines of delivered code [21]. Defect Density is used here to measure the quality of the software product. It gives an indication of quality improvement achievements in successive releases of certain software. The lower the number of defect density, the better the software quality is. Defect density can be computed using equation 1 as follows:

$$\text{Defect Density} = \frac{\text{Number of Defects}}{\text{KLOC}} \tag{1}$$

Defect density is correlated with number of developers and software size jointly [22]. similar results are obtained in [21], where projects size is found to be an affecting factor (large projects are found to have lower defect density). Development mode is found to be another factor that affects defect density rate (open source projects are found to have a lower defect density).

4 Data Collection

We conducted the empirical study on four open source systems. In selecting the subjected systems, we used several criteria. First, we want well-known systems that are used very widely. Second, systems had to be sizable, so we can understand the issues that appear in the evolution of realistic, multi-developer software. Third, the systems had to be actively maintained. Finally, the data of these systems had to be publicly available. Public availability of the data used for empirical studies is crucial. A theory of software evolution must be based on empirical results, verifiable and repeatable [5]. Characteristics of the selected software systems are listed in Table 1. An overview of each system is provided in the following paragraphs.

Table 1. Selected Software Systems

System	Versions	LOC
Camel	1.0-1.6	3594-113055
jEdit	4.0-4.3	144803-202363
POI	1.5-3.0	55428-129327
Xerces	1.0-1.4	90718-141180

Apache Camel is a powerful open source integration framework based on known Enterprise Integration Patterns with powerful Bean Integration. jEdit is a mature programmer' s text editor with hundreds (counting the time developing plugins) of person-years of development behind it. It is written in Java and runs on any operating system with Java support, including Windows, Linux, Mac OS X, and BSD. The POI project consists of APIs for manipulating various file formats based upon Microsoft' s OLE 2 Compound Document format, and Office OpenXML format, using pure Java. Xerces is a parser that supports the XML 1.0 recommendation and contains advanced parser functionality, such as support for XML Schema 1.0, DOM level 2 and SAX version 2.

The data for this study were collected by [19] and are available online at the PROMISE repository. This data was widely used in the software engineering literature for different purposes [23, 24, 25]. The collected measures' data for the four systems are added up correspondingly into one data set along with the relevant values for coupling, cohesion, and complexity measures. Descriptive statistics (Min, Max, Median, Std. dev.) defined the minimum, maximum, median, and standard deviation measures. Table 2 shows descriptive statistics about the selected measures.

Table 2. Descriptive Statistics of the Measures

Measures	**Min**	**Max**	**Med**	σ
CBO	0	187	9.8	15.20
RFC	0	498	27.72	39.8
Ca	0	184	5.33	13.79
Ce	0	95	5.24	6.75
CBM	0	25	1.59	3
LCOM	0	41719	116.3	933.3
LCOM3	0	2	1.14	0.67
CAM	0	1	0.47	0.25
WMC	0	409	10	18.8
Avg(CC)	0	25.14	1.28	1.3

5 Data Analysis and Results

In order to answer the first question, we need to know which aspects of coupling is measured by any of the chosen coupling measures. Same thing holds for cohesion. To achieve that, we use the well-known Principal Component Analysis (PCA) which is a standard statistical procedure that uses orthogonal transformation to identify the underlying, orthogonal dimensions that explain relations between the variables in the data set. We conducted the experiments using the R statistical software (version 3.1.1) and we used R' s Procomp procedure to our data to produce principal components. The analysis is done on the entire data set of the considered measures.

The objectives of principal component analysis are to discover or reduce the dimensionality of the data set and identify new meaningful underlying variables. PCA is a de facto technique for uncovering the underlying orthogonal dimension that explains variables relations in a dataset. PCA is used in our case to identify measures (i.e, groups of independent variables) that measure the same underlying dimension (i.e., mechanism that defines coupling and cohesion among classes). Principal Components (PCs) are linear combinations of independent variables. The number of PCs is less than or equal to the number of original variables. PCs are interpreted as follows. Each new PC is orthogonal to all previously calculated PCs and captures a maximum variance under these conditions.

5.1 Coupling Evolution Analysis & Results

In this section we apply the PCA approach to the coupling measures to specify any correlations among them. If a group of coupling measures are strongly correlated, these measure are likely to measure the same underlying dimension (i.e., class property) of the object to be measured.

Table 3. Rotated Components of Coupling measures

	PC1	**PC2**	**PC3**	**PC4**
Proportion	39%	22%	20%	19%
Cumulative	38%	60%	80%	100%
CBO	0.92	0.34	-0.02	0.20
RFC	0.21	0.39	0.11	0.89
Ca	0.99	0.00	-0.03	0.10
Ce	0.18	0.91	0.04	0.37
CBM	-0.03	0.04	1.00	0.08

By analyzing the coefficients associated with every coupling measure within each rotated component given in Table 3, we interpret the identified PCs as the following:

- PC1 (39%): CBO and Ca measures count inbound coupling through method invocations. The correlation betweeen the two measures is high. We can use one of them rather than using both. Apparently, the afferent couplings measure is the contributing measure as it has higher PC value.
- PC2 (22%): Ce captures outbound coupling through method invocations.
- PC3 (20%): CBM captures coupling between inherited and redefined methods.
- PC4 (19%): RFC counts the number of accessible methods.

5.2 Cohesion Evolution Analysis & Results

We also conducted PCA analysis on the selected cohesion measures. We want to see if any correlations exists between these measures.

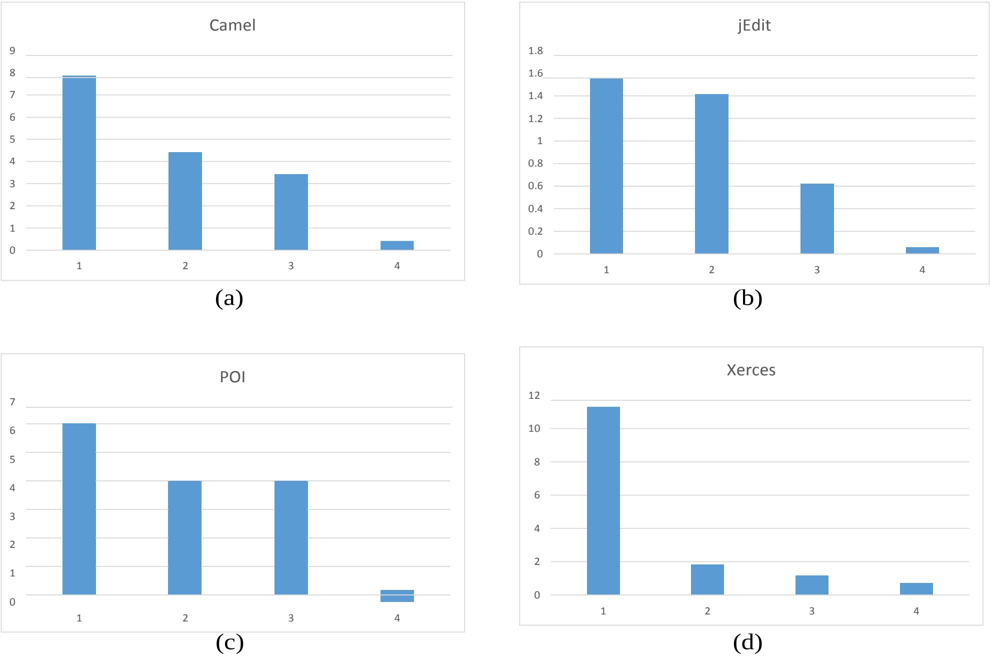
Table 4. Rotated Components of Cohesion measures

	PC1	**PC2**	**PC3**
Proportion	33%	33%	33%
Cumulative	33%	67%	100%
LCOM	1	-0.01	-0.07
LCOM3	-0.01	0.98	0.20
CAM	-0.08	0.21	0.98

By analyzing the coefficients associated with every cohesion measure within each rotated component given in Table 4, we found that the identified PCs as each on of these cohesion measures is unique and does not overlap with the others.

5.3 Defect DensityEvolution

We measured the defect density of each version of the adopted open source software in order to investigate if the evolution of each software reduces the defect density or increases it over the different releases. Figure 1 shows how defect density evolved in the selected systems. The defect density is shown to improve for each of the tested software over the successive releases.



(a)

(b)

(c)

(d)

Figure 1. The Evolution of Defect Density in the Selected Systems.

5.4 Discussion

According to our PCA analysis, the coupling measures that can be used to measure the system' s modularity are Ca, Ce, CBM and RFC. The CBO measure has been excluded as the Ca measures the same dimension. The cohesion measures that can be used to measure the system' s modularity are LCOM, LCOM3, and CAM. These measures along with the complexity measures WMC and CC measures altogether are the set of measures that measure the Modularity of a system or software program. This answers the first research question.

Table 5 shows the coupling, cohesion, and complexity evolution of the four selected systems. There are three notions which characterize good and bad things about modules, coupling (we want low coupling between modules), cohesion (we want highly cohesive modules), and complexity (we want modules that have low complexity) [2]. Modularity is a concept in which a software is decomposed of several distinct and logically cohesive sub-units, offering services through a well-defined interface [13]. Excessive inter-module

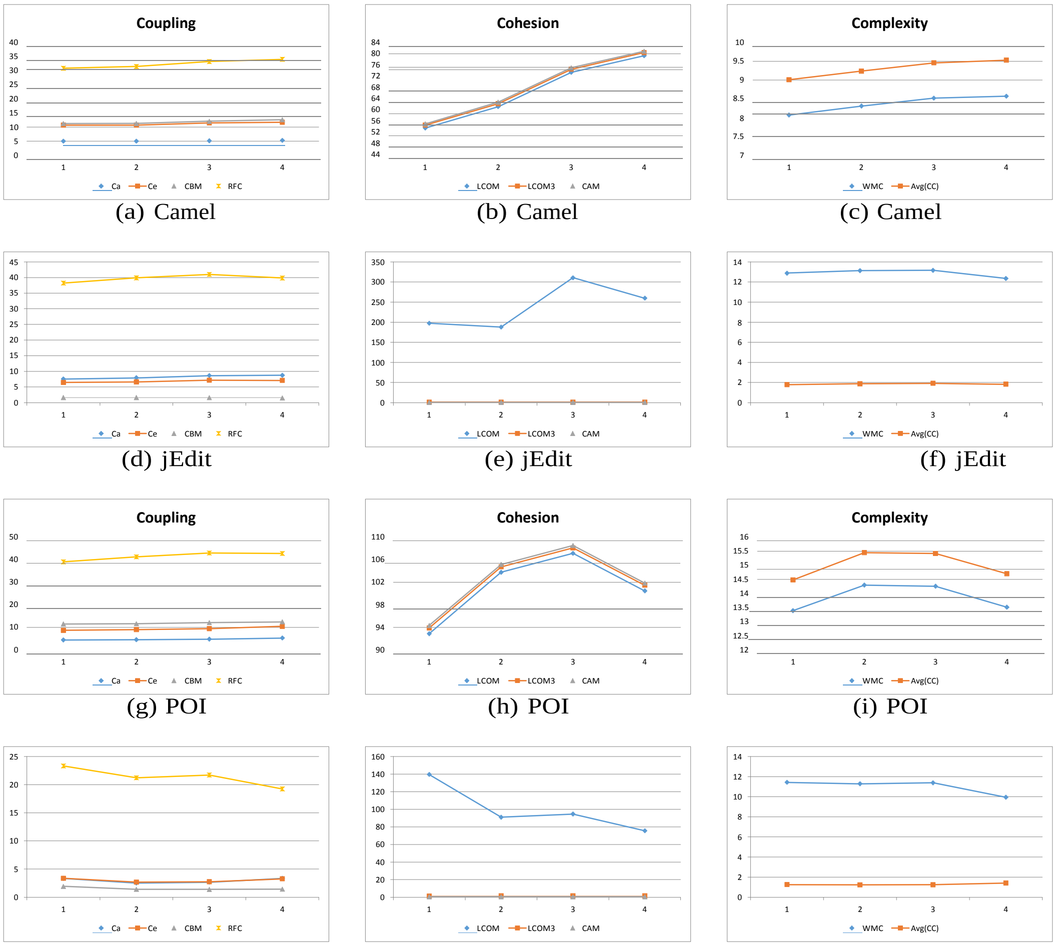
Table 5. Modularity Evolution of the Selected Systems

			Ver. 1	Ver. 2	Ver. 3	Ver. 4
Camel	Coupling	Ca	4.99	5.02	5.11	5.27
		Ce	5.69	5.62	6.33	6.43
		CBM	0.56	0.64	0.61	0.91
		RFC	19.63	20.23	21.2	21.42
	Cohesion	LCOM	53.65	61.24	73.42	79.33
		LCOM3	0.99	1.08	1.11	1.1
		CAM	0.48	0.5	0.49	0.49
	Complexity	WMC	8.07	8.31	8.52	8.57
		Avg(CC)	0.94	0.93	0.94	0.96
jEdit	Coupling	Ca	7.51	7.93	8.62	8.74
		Ce	6.43	6.63	7.16	7.1
		CBM	1.61	1.59	1.55	1.5
		RFC	38.24	39.87	40.98	39.85
	Cohesion	LCOM	197.38	187.89	310.76	259.91
		LCOM3	1.05	1	0.99	1.09
		CAM	0.47	0.45	0.44	0.46
	Complexity	WMC	12.88	13.13	13.16	12.35
		Avg(CC)	1.79	1.87	1.92	1.83
POI	Coupling	Ca	4.36	4.51	4.7	5.23
		Ce	4.31	4.48	4.68	5.22
		CBM	2.78	2.62	2.7	1.95
		RFC	27.56	29.65	30.9	30.35
	Cohesion	LCOM	92.87	103.76	107.12	100.46
		LCOM3	1.02	0.97	0.98	1
		CAM	0.44	0.42	0.43	0.38
	Complexity	WMC	13.39	14.3	14.26	13.51
		Avg(CC)	1.09	1.15	1.16	1.19
Xerces	Coupling	Ca	3.33	2.52	2.67	3.35
		Ce	3.38	2.68	2.75	3.27
		CBM	1.93	1.41	1.38	1.43
		RFC	23.33	21.23	21.7	19.24
	Cohesion	LCOM	139.48	91	94.52	75.49
		LCOM3	1.22	1.49	1.47	1.47
		CAM	0.52	0.51	0.5	0.52
	Complexity	WMC	11.43	11.28	11.38	9.94
		Avg(CC)	1.26	1.22	1.24	1.4

dependencies has been acknowledged to be an indicator of poor design and decrease the comprehending of components in isolation [26].

Figure 2 shows the evolution of coupling, cohesion, complexity measures of the selected systems over four different releases for each system. The X-axis represents the release number while the Y-axis represents the measures data. As can be seen from figure 2 a, d and g, we can see that there is a minor change in the Ca, Ce and CBM coupling measures. But there is slightly more increase in the RFC measure in Camel, jEdit and POI while the RFC slightly decrease for Xerces. Hence, coupling is slightly increasing while the software is evolving, this indicates that the modularity is not improving over time.

Figure 2 b show that the LCOM and LCOM3 are increasing over the various releases of Camel. Figure 2 e shows that for jEdit, the cohesion level indicated some improvement in the second release but lost it in the third release then again made some progress in the forth release but still not as good as in the first release which meanins that overall the



(a) Camel (b) Camel (c) Camel

(d) jEdit (e) jEdit (f) jEdit

(g) POI (h) POI (i) POI

(j) Xerces (k) Xerces (l) Xerces

Figure 2. Modularity Evolution of the Selected Systems.

cohesion is not improving significantly in jEdit. Figure 2 h shows cohesion of the POI software. overall the cohesion measures indicate that the cohesion is not improving till the third release, then the cohesion started improving in the fourth release but still not good as in the first release. Xerces is in better situation that the other software where the LCOM is improving over the various successive releases, LCOM3 and CAM are kept in a steady level. Accordingly, we can notice that the cohesion measures shows that the Camel jEdit and POI software are not improving while evolving.

Regarding the complexity measure, Figure 1 c shows that the Camel software complexity is increasing over the various releases. This indicates that there is no restructuring activities is done in these four versions. jEdit software shows some improvement over its successive releases. This means that some restructuring activities have took place but did not significantly improve the jEdit complexity. POI software complexity has increased in the second release, but started to decrease in the following releases but still the complexity is slightly more than that of the first release. For Xerces software complexity shows a noticeably improvement in WMC measure and some improvement in the avg(CC) measure in the fourth

Table 6. Spearman correlation coefficient of Modularity measures and Bugs

	Ca	Ce	CBM	RFC	LCOM	LCOM3	CAM	WMC	avg(CC)
Bugs	0.14	.19	.12	.38	0.40	0.22	0.15	0.32	0.36

release.

Accordingly, the various measures of coupling, cohesion and complexity of Camel jEdit and POI software show that the modularity of three of the software is not improving overall! Xerces software is in a better situation where its measures showed some improvements. This means that modularity is not improving significantly, hence we can not say that an effective restructuring has took place, although defect density is improved. This means that designers and developers where concerned with solving bugs and problems without paying enough attention to restructuring that aims to improve software structures quality. Hence, we believe that restructuring is needed in the coming releases to improve software quality. This answers the second research question.

Moreover, to test the relationship between the modularity measures and the number of bugs in a software version, we have conducted a correlation analysis. Correlation analysis studies the degree to which changes in the value of an attribute (one of the modularity measures) are associated with changes in another attribute (number of faults in a version). The Spearman correlation is preferred instead of Pearson correlation because the former ignores any assumptions about the data distribution [27].

Ifthe measuretends to increase whenthe number of bugs increases, the Spearman correlation coefficient is positive. Ifthe measuretends to decrease whenthe number of faultsincreases, the Spearman correlation coefficient is negative. Table 6 shows that RFC, LCOM, WMC, and Avg CC have a moderate correlation with number of faults. These results are very similar to Johari and Kaur study [28]. Accordingly, our data shows that there is a moderate relation between modularity measures and number of faults in our software sample.

6 Threats to Validity

The conducted research in this paper is exposed to possible validity threats which are defined and discussed in [29]:

- Construct Validity: The various measures we used (coupling measures, cohesion measures, complexity measures, defect measures and correlation measures) are well documented in literature. The data are collected for four open source software which are public ally available.
- External Validity: Our data set is collected from the software engineering repository PROMISE. We have collected data for four open source software over four successive releases for each. Results obtained based on this data set should be relevant and valid for other releases of the studied software as well as other ones.
- Internal Validity: All the needed data pieces in this study have been collected by the researchers from the mentioned data repository. missing data could be there but has minimal effect of the conducted analysis and conclusion.
- Conclusion Validity: Our analysis have been conducted based on the collected data

set. A threat to the conclusion validity is can be related to how the data is reported in the repository; e.g what is considered a fault or bug for example when a certain incident occur in the system would it be considered as a fault or change request. As we are talking about open source software, the developers and designers skills participating in the four project from different locations with different experience skills may form a threat. The number of projects used in this research may also form a kind of threats. We have used data sets for four software projects. Although we believe the results can be generalized for other open source projects, enlarging the data set by adding more projects may provide more reliable results.

7 Related Work

Open-source systems are usually developed by distributed teams, without frequently meeting face-to-face, and communicating only by electronic means. Achieving high modularity in open source allows multiple developers to work on the same software entity without issues [15]. This new structure is totally unlike the common software engineering practices during the times of Lehman’ s software evolution laws [5]. Lehman et al. have built the well-known research on the evolution of large software systems. Lehman’ s laws are based on case studies of several large software systems, suggest that as systems grow in size, it becomes increasingly difficult to add new code unless clear steps are taken to restructure the overall design.

MacCormack et al. [30] employed Design Structure Matrix (DSM) to compare and contrast the design structures of two software systems, Linux kernel and Mozilla web browser. They used a clustering algorithm to measure dependencies by different parts of the system and calculated marginal changes in cost rather than the total cost of the matrix. However, the comparison between these two systems critically depends on selecting versions of the systems that are comparable in terms of number of source files. One motivation of our work was to remove this restriction, and to allow the comparison of code bases of different size. LaMantia et al. [31] examined the evolution over time of two software systems, Apache Tomcat and another closed source server product. They introduced a rough measure that mimics the change ratio between the consecutive versions in the software evolution. The authors concluded that DSM could, to some extent, explain how modularization allow for different rates of evolution to occur in different modules.

Koch found differences in the evolution of open-source software projects of different sizes [32]. He found that small open-source software projects fulfill some of the laws. However, large software projects do not follow them at all. These projects have a large number of participants and an unbalanced workload among participants. One of the essential characteristics of software systems is evolution. Several research studies aimed at explaining and understanding the evolution in open source software projects. Breivold et al. [33] conducted a systematic literature review of enormous studies, which investigated the evolution of open source software systems. Another direction has emphasized how software measures can be applied to software evolution [34] where they provided ways in how software measures have been and can be used to analyze software evolution. They suggested that measures are good candidates to understand the quality evolution of a software system by considering its successive releases. Particularly, measures can be used to measure whether the quality of a software has improved or degraded between two releases. Lee et al. [35]

provided a case study of one open source software, JFreeChart, evolution with software measures. They studied the evolution in terms of size, coupling and cohesion, and discussed its quality change based on the Lehman’ s laws of evolution [5]. Neamtiu et al. [36] conducted an empirical analysis on the evolution of nine popular open-source programs and investigated Lehman’ s evolution laws where their study confirmed that continuing change and continuing growth are still applicable to the evolution of today’ s open-source software.

Neamtiu et al. [36] used source code measures with project defect information to analyze software growth, characterize software changes, and assess software quality. Murgia et al. [37] focused their study on software quality evolution in open source projects using agile practices. They used several OO measures to study how bug distribution relates to software evolution. They found that there is no a single metric that is able to explain the bug distribution during the systems evolution. Eski et al. [38] investigated the relationship between OO measures and changes in open source software systems and proposed a metric-based approach to predict change-prone classes.

Other researchers have conducted similar empirical studies and prosed some new metrics, for instance Li et. al. [39] studied the evolution of an object-oriented system using the OO metrics suggested by Chidamber and Kemerer to measure the class-level design and proposed three new metrics to study OO system evolution (System Design Instability (SDI), Class Implementation Instability (CII), and System Implementation Instability (SII)). Drowin [40] analyzed empirically the quality evolution of an open source software using a control flow based metric (Quality Assurance Indicator - Qi) which they claimed that Qi metric reflects properly the quality evolution of the system.

In this paper, we study the modularity evolution of four open-source systems. The focus of this study is not the Lehman’ s Law but the modularity using coupling, cohesion, and complexity measures.

8 Conclusion

Enhancing our ability to understand and capture software evolution is essential for better software quality and easier software maintenance process. one of the vital features that reflects the software quality is its structures quality. structures quality has relationship with software modularity. We have used modularity measures to give indications about software structures quality. In this research work, we have used empirical data related to four OO open source programs to answer two main research questions namely: what measures can be used to measure the modularity level of software and secondly, did the modularity level for the selected open source software and their structures quality improve over time? By investigating the modularity measures as mentioned in the SQuaRE standard and various other coupling and cohesion measures, we have identified the main measures that can be used to measure software modularity. Based on our analysis, the modularity of these four systems did not show a significant improvement in their modularity and structures quality over time. However, the defect density is improving over time.

References

[1] Mamdouh Alenezi and Mohammad Zarour. Modularity measurement and evolution in object-oriented open-source projects. In *International Conference on Engineering & MIS 2015 (ICEMIS’15)*. ACM, 2015.

[2]Frank F Tsui. *Essentials of software engineering*. Jones & Bartlett Publishers, 2013.

[3]Edward E. Ogheneovo. Development of a software maintenance cost estimation model: 4th gl perspective. *Journal of Computer and Communications*, 2:1–16, 2014.

[4]Ana Filipa Nogueira. Predicting software complexity by means of evolutionary testing. In *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, pages 402–405. ACM, 2012.

[5]Michael W Godfrey and Daniel M German. On the evolution of lehman’s laws. *Journal of Software: Evolution and Process*, 26:613619, 2013.

[6]Segla Kpodjedo, Filippo Ricca, Philippe Galinier, Giuliano Antoniol, and Yann-Ga¨el Gu´eh´eneuc. Studying software evolution of large object-oriented software systems using an etgm algorithm. *Journal of Software: Evolution and Process*, 25(2):139–163, 2013.

[7]Narasimhaiah Gorla and Ravi Ramakrishnan. Effect of software structure attributes on software development productivity. *Journal of Systems and Software*, 36(2):191–199, 1997.

[8]Sunny Huynh, Yuanfang Cai, Yuanyuan Song, and Kevin Sullivan. Automatic modularity conformance checking. In *ACM/IEEE 30th International Conference on Software Engineering, 2008. ICSE’08.*, pages 411–420. IEEE, 2008.

[9]David Lorge Parnas. On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, 15(12):1053–1058, 1972.

[10]Sunny Wong, Yuanfang Cai, Miryung Kim, and Michael Dalton. Detecting software modularity violations. In *Proceedings of the 33rd International Conference on Software Engineering*, pages 411–420. ACM, 2011.

[11]Alessandro Rossi and Alessandro Narduzzo. Modular design and the development of complex artifact lesson fron free open source software. Technical report, Department of Computer and Management Sciences, University of Trento, Italy, 2003.

[12]ISO/IEC. Systems and software engineering - systems and software quality requirements and evaluation (square). *ISO/IEC 25010 - System and software quality models*, 2011.

[13]Carliss Young Baldwin and Kim B Clark. *Design rules: The power of modularity*, volume 1. MIT press, 2000.

[14]Grady Booch, Robert A Maksimchuk, Michael W Engel, Bobbi J Young, Jim Conallen, and Kelli A Houston. *Object-oriented analysis and design with applications*, volume 3. Addison-Wesley, 2008.

[15]Mark Aberdour. Achieving quality in open-source software. *IEEE Software*, 24(1):58–64, 2007.

[16]Juliana de AG Saraiva, Micael S de Fran¸ca, S´ergio CB Soares, JCL Fernando Filho, and Renata MCR de Souza. Classifying metrics for assessing object-oriented software maintainability: A family of metrics’ catalogs. *Journal of Systems and Software*, 103:85–101, 2015.

[17]Mourad Badri, Linda Badri, and Fadel Tour´e. Empirical analysis of object-oriented design metrics: Towards a new metric using control flow paths and probabilities. *Journal of Object Technology*, 8(6):123–142, 2009.

[18]Mamdouh Alenezi and Khaled Almustafa. Empirical analysis of the complexity evolution in open-source software systems. *International Journal of Hybrid Information Technology*, 8(2):257–266, 2015.

[19]Marian Jureczko and Diomidis Spinellis. Using object-oriented design metrics to predict software defects. *Models and Methods of System Dependability. Oficyna Wydawnicza Politechniki Wroc~lawskiej*, pages 69–81, 2010.

[20]Shyam R Chidamber and Chris F Kemerer. A metrics suite for object oriented design. *IEEE Transactions on Software Engineering*, 20(6):476–493, 1994.

[21]Syed Muhammad Ali Shah, Maurizio Morisio, and Marco Torchiano. An overview of software defect density: A scoping study. In *19th Asia-Pacific Software Engineering Conference (APSEC)*, volume 1, pages 406–415. IEEE, 2012.

[22]Cobra Rahmani and Deepak Khazanchi. A study on defect density of open source software. In *IEEE/ACIS 9th International Conference on Computer and Information Science (ICIS)*, pages 679–683. IEEE, 2010.

[23]Giuseppe Scanniello, Carmine Gravino, Andrian Marcus, and Tim Menzies. Class level fault prediction using software clustering. In *IEEE/ACM 28th International Conference on Automated Software Engineering (ASE)*, pages 640–645. IEEE, 2013.

[24]Burak Turhan, Ay¸se Tosun Mısırlı, and Ay¸se Bener. Empirical evaluation of the effects of mixed project data on learning defect predictors. *Information and Software Technology*, 55(6):1101–1118, 2013.

[25]Mamdouh Alenezi, Shadi Banitaan, and Qasem Obeidat. Fault-proneness of open source systems: An empirical analysis. In *International Arab Conference on Information Technology (ACIT2014)*, pages 164–169, 2014.

[26]Hongyu Pei Breivold, Ivica Crnkovic, and Magnus Larsson. Software architecture evolution through evolvability analysis. *Journal of Systems and Software*, 85(11):2574–2592, 2012.

[27]Jay Devore. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 2015.

[28]Kalpana Johari and Arvinder Kaur. Validation of object oriented metrics using open source software system: an empirical study. *ACM SIGSOFT Software Engineering Notes*, 37(1):1–4, 2012.

[29]Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[30]Alan MacCormack, John Rusnak, and Carliss Y Baldwin. Exploring the structure of complex software designs: An empirical study of open source and proprietary code. *Management Science*, 52(7):1015–1030, 2006.

[31]Matthew J LaMantia, Yuanfang Cai, Alan D MacCormack, and John Rusnak. Analyzing the evolution of large-scale software systems using design structure matrices and design rule theory: Two exploratory cases. In *Seventh Working IEEE/IFIP Conference on Software Architecture*, pages 83–92. IEEE, 2008.

[32]Stefan Koch. Software evolution in open source projectsa large-scale investigation. *Journal of Software Maintenance and Evolution: Research and Practice*, 19(6):361–382, 2007.

[33]Hongyu Pei Breivold, Muhammad Aufeef Chauhan, and Muhammad Ali Babar. A systematic review of studies of open source software evolution. In *17th Asia Pacific Software Engineering Conference (APSEC), 2010*, pages 356–365. IEEE, 2010.

[34]Tom Mens and Serge Demeyer. Future trends in software evolution metrics. In *Proceedings of the 4th international workshop on Principles of software evolution*, pages 83–86. ACM, 2001.

[35]Young Lee, Jeong Yang, and Kai H Chang. Metrics and evolution in open source software. In *Seventh International Conference on Quality Software, 2007. QSIC’07.*, pages 191–197. IEEE, 2007.

[36]Iuan Neamtiu, Guowu Xie, and Jianbo Chen. Towards a better understanding of software evolution: an empirical study on open-source software. *Journal of Software: Evolution and Process*, 25(3):193–218, 2013.

[37]Alessandro Murgia, Giulio Concas, Roberto Tonelli, and Ivana Turnu. Empirical study of software quality evolution in open source projects using agile practices. In *Proc. of the 1st International Symposium on Emerging Trends in Software Metrics*, page 11, 2009.

[38]Sinan Eski and Feza Buzluca. An empirical study on object-oriented metrics and software evolution in order to reduce testing costs by predicting change-prone classes. In *IEEE Fourth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pages 566–571. IEEE, 2011.

[39]Wei Li, L Etzkorn, C Davis, and J Talburt. An empirical study of object-oriented system evolution. *Information and Software Technology*, 42(6):373–381, 2000.

[40]Nicholas Drouin, Mourad Badri, and Fadel Tour´e. Metrics and software quality evolution: A case study on open source software. In *Proceedings of the 5th International Conference on Computer Science and Information Technology, Hong Kong*, 2012.

Intrusion Detection – A Text Mining Based Approach

Gunupudi RajeshKumar[1], N Mangathayaru[2], G Narsimha[3]

[1,2]Faculty of Information Technology, VNR VJIET, India
[3]Faculty of Computer Science and Engineering, JNT University, Jagityal, India

Abstract: Intrusion Detection is one of major threats for organization. The approach of intrusion detection using text processing has been one of research interests which is gaining significant importance from researchers. In text mining based approach for intrusion detection, system calls serve as source for mining and predicting possibility of intrusion or attack. When an application runs, there might be several system calls which are initiated in the background. These system calls form the strong basis and the deciding factor for intrusion detection. In this paper, we mainly discuss the approach for intrusion detection by designing a distance measure which is designed by taking into consideration the conventional Gaussian function and modified to suit the need for similarity function. A Framework for intrusion detection is also discussed as part of this research.

Keywords: system calls, intrusion, prediction, classification, kernel measures

1. Introduction

Intrusion may be defined as an activity that essentially attempts to compromise integrity, authenticity, confidentiality, availability of system resources. For an organization to be safe, efficient and most reliable it must maintain several layers of security. These include network security and information security. This challenge is becoming much more complex currently as systems and services are becoming complex facilitating several new possibilities for attackers.

One may achieve information security by maintaining confidentiality, integrity and availability. Also as the data is enormously increasing and turning into big data, various design challenges, data analysis challenges, requirement for new algorithms, methodologies and measures are coining out. This further makes the situation more complex to handle.

Another problem which coins is the curse of dimensionality. Intrusion detection is not free from the problem of dimensionality and must be handled without fail for accurate results. The process called knowledge discovery from databases may be used in hand with the methods and methodologies of intrusion detection. Data Mining and Intrusion Detection go hand in hand now days.

Intrusion detection systems combine data mining methods, methodologies and algorithms along with the attack detection in to the system so that , the system can detect the intrusion dynamically. Similarly, the Intrusion detection systems (IDS) which mainly use anomaly detection mechanisms try to discover the abnormal behaviours.

In spite of several detection mechanisms available, there is a dearth of proper mechanism which can fix the behaviour as the intrusion. Though the findings and methods work for standard datasets but they ultimately fail over the real time datasets generated dynamically.

Figure 1 shows, the framework of NIDES intrusion detection system. We can also combine both the signature based and anomaly based intrusion techniques to forma hydrid technique to come up with the decision on normal and intrusive traffics. The figure 2 shows the knowledge discovery framework depicting each stage of the knowledge discovery process which may be used along with intrusion detection mechanisms to improve the efficiency and optimize the output results.

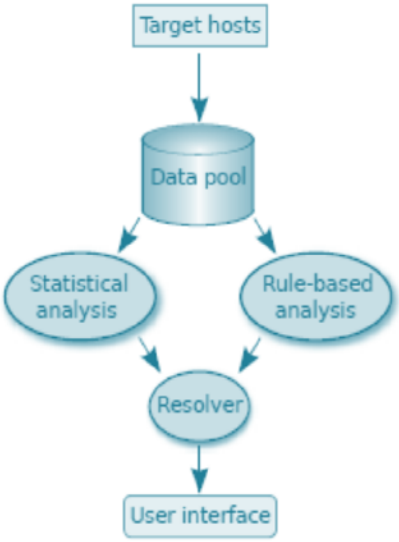


Fig 1: Framework of NIDES Intrusion Detection

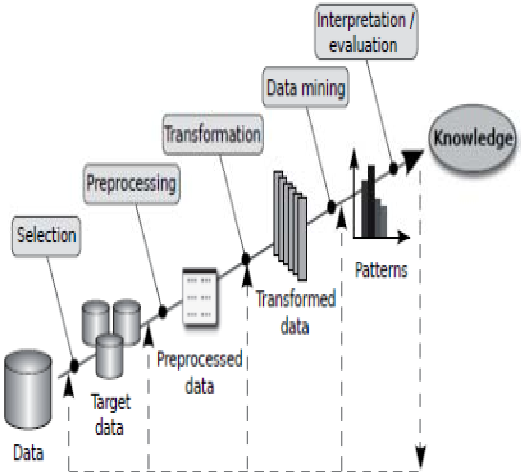


Fig 2 : Framework of KDD process

Securing internal and external resources for any organization in an unauthorized way, is becoming a major alarming problem now-a-days. Any sensitive information usually attracts attention from intruders leading resources to become vulnerable. The design of intrusion detection systems mainly functions based on two concepts. The former is the misuse based intrusion detection system and the later is the anomaly based intrusion detection system [26]. For the misuse based intrusion detection system, it is mandatory to construct a knowledge base which is useful in taking decisions whether the incoming request is normal or intrusion. The knowledge base is a collection of known signatures and instances of intrusion attacks.

Whenever a new request arrives, its signature will be cross checked with the already existing signatures in knowledge base and the decision is taken. Alert will be generated by the intrusion detection system, in case if it is a threat.

The second methodology is anomaly based, where the intrusion detection system learns the behavior of the system, and will immediately generate an alert in the case of deviation from the normal behavior [1,2,5].

Research contribution towards building IDS using different data mining techniques has been extensively studied in literature. In order to detect threats, it is advantageous to use soft computing techniques rather than traditional approaches for construction of the IDS.

2. Related Work

Intrusion detection system monitors all the incoming traffic and restricts entry of an unauthorized attempt to protect the resources by applying suitable rules. Recent research publications in intrusion detection algorithms concentrated more on the feature extraction from the data. The following is the list of various related techniques published in various journals for intrusion detection as shown in Fig. 1.

2.1 Text Processing

In this approach for intrusion detection, the system calls serve as the source for mining and predicting any chance of intrusion. When an application runs, there might be several system calls which are initiated in the background. These system calls form the basis and the deciding factor for intrusion detection [1, 2, and 3]. The approach of intrusion detection using text processing has been one of the research interests among researchers working in the area of network and information security.

This technique uses system call sequences [1] by applying text processing techniques. Alok Sharma et. al 2007, in his paper discussed the intrusion detection mechanism using text processing technique k-nearest neighbor (kNN) classifier. This classifier was used on the DARPA 1998 database and the results were proved to be better than other algorithms[2]. In their paper, Alok Sharma et. al 2007, demonstrated the cosine similarity measure and binary similarity measure.

For the first time Liao et. al[2] used the cosine similarity measure and later Rawat et. al. [3] extended by introducing the weight component for the system calls.

Alok Sharma et. al. proposed a new similarity measure which not only considers the frequency of the system calls rather than the number of common system calls between the processes.

2.2 SVM

SVM is one of the finest supervised methods used for classification. It includes learning algorithms through which, the training data gets classified. In order to get more quality in the classification process, SVM uses high dimension values for classification [12].

Initially, the training values are given through which these values are classified. SVM is generally used for the classification and Regression. In the process of determining the intrusion using system calls, if the number of system calls is too many, it will be difficult to perform the classification keeping performance unchanged. In order to have the performance of classification process unchanged, the dimensions need to be reduced without affecting the quality

2.3 Signature Detection

Intrusions are generally detected by matching captured pattern with already preconfigured knowledge base. The rate of false alarms in the case of unknown attacks is very high [25]. In his paper ,Yuxin Meng at.al. narrates that the Intrusion Detection System based on Signature [14] smells an attack by analyzing its stored signatures with information within the packet payloads. The signature may be a collection of rules which can be formed as an identity, which shall be stored in the database. In conclusion, the signature based intrusion detection is one of the prominent approaches to detect the threats, even though it has a drawback of possibility of occurrence of false alarms. The detection accuracy is high in the case of previously known attacks and computational cost is very less.

2.4 Genetic Algorithms

These GA techniques are generally used to select the best features that are used for IDS and when compared to other methods, has better efficiency. This is an approach which is, slightly trickery, complex and hence need to be used in specific manner rather in general approach [9, 23]. Studies based on both Genetic Algorithms and Fuzzy rule based systems can be classified as Michigan, Pittsburgh approaches. The Pittsburgh method uses a set of if-then fuzzy rules are coded as an individual. In Michigan only a single if-then fuzzy rule is coded as individual.

2.5 Fuzzy Logic Approach

The fuzzy logic approach considers the approximate theory rather than taking exact inference from predicate logic into consideration. These methods use quantitative features. This technique provides improved flexibility to some uncertain problems. However, when compared to the artificial neural network approach, the detection accuracy is lesser [25]. These fuzzy approaches can also be used for the anomaly detection as the features can also be considered as fuzzy variables [22]. As long as the observation lies within prescribed intervals, this kind of processing schemes are to be considered as normal (Dickerson, 2000)[23]. Sometimes intrusion detection systems based on the anomaly flags observed, activities that keep deviating from normal attribute patterns [20].

2.6 Anomaly Based Approach

These approaches use statistical test on the collected behavior to identify the occurrence of intrusion. Time taken to identify is more in this method. The detection accuracy is directly proportional to the collected amount of behavioral pattern features. The rate of false alarms is less in the case of unknown attacks [25]. Sang Hyun Oh, et. al. (2003) proposed in his paper [4]that an anomaly detection measure that uses a clustering algorithm which models the normal behaviors of users activities.

The statistical analysis predictive pattern generation and data mining techniques classify an anomaly detection model [5]. The values related to features of user activity represent the corresponding feature rate for the execution of the activity. Because of this, the value domain of the features is generally mentioned in the form of integers, makes association and sequence data mining inapplicable.

Hence only frequent item sets be found among the finite number of categorical items. In the contrary, this kind of problems can be better handled by the clustering as it is purely based on the similarity of data.

For analyzing, the common properties of all transactions of a user, anomaly based intrusion detection techniques are included in a host based IDS. The DBSCAN, JAM, ADMIT and EMERALD are few algorithms combined for Intrusion Detection Systems for ensuring successful threat detection. The anomaly based intrusion detection algorithms can be classified into three categories [21] as follows:

i. Statistical based : Nature of the process involved, is univariate, multivariate, time series model.

ii. Knowledge based: Nature of the process involved are FSM, Decryption languages, Expert systems.

iii. Machine Learning based: All soft computing techniques comes into this group

2.7 Association Rule Based

These techniques are used for only known attach signatures and/or relevant attacks in the misuse detection. Total unknown attacks are not at all detected and moreover it requires more number of database scans, to generate the rule base [25]. Lee, at.al. initiated the concept of using association rules for intruder detection solutions and was extended in [15, 16, 17].

2.8 Dimensionality Reduction

In order to avoid false alarms, two techniques need to be completed without fail. They are Preprocessing techniques and Dimensionality Reduction techniques. All the system calls that were captured by the Intrusion Detection System, many not play a role in deciding whether the incoming request is a threat or not. In this case, the system calls which were not important or irrelevant to the threat detection process must be identified and removed from the database. That is how a preprocessing system will decide each unit of data, whether it is a normal data or an anomalous data.

The preprocessing tasks involve activities like dataset creation, cleaning, integration, feature building. These most important steps to be discussed are:

1. Dataset Creation: Data to be identified and collected in order to proceed for the preprocessing process. The data need to be separately identified for Training Phase and Testing Phase.

2. Feature Building: To improve the discriminative properties for the anomaly detection process, additional features are described for the data. This feature building can be done manually or using some automated tools.

3. Knowledge Discovery Approach

Most of the significant works carried for finding intrusion detection may be classified into the following classes

1. Machine Learning Based Approach
2. Unsupervised Learning Based Approach
3. Supervised Learning Based Approach
4. Genetic programming Based Approach

3.1 Machine Learning Approach

Machine learning is a self learning approach which requires a formal system which can update itself continuously each time the new data is generated and added to the system. In essence, it must be an autonomous system which can address the continuous changes coined out and integrate the knowledge database.

This process requires ability to learn from experience, analytical capability, self learning capability, ability to handle dynamic changes so as to be self updated.

In essence, the major task in the machine learning algorithms is to design, analyze, develop, and implement various algorithms and methodologies which guide the machines (computer systems) to gain the self learning capability. Machine learning may be classified into supervised and unsupervised learning techniques [20].

3.2 Supervised Learning Approach

In this approach for intrusion detection, we must know the class label to build the knowledge database or knowledge rules. This is because of this reason; we call it as supervised learning technique or classification.

Given a dataset, we split the dataset into training and testing sets and build the knowledge using the training set. Then we use the samples from the testing set to test the class label of the test case chosen from the testing test. In short, the task of supervised learning is to build a classifier which can effectively approximate the mapping between input and output samples of training.

Once we build a classifier, this is followed by measuring the classification accuracy. Classification requires choosing an appropriate function which can estimate the class label. This is followed by measuring the classification accuracy.

The most popular classifiers include Decision tree based Classifier; ANN based classifier, KNN Classifier, SVM Classifier. The simplest non-parameter classifier is the KNN-classifier which is used to estimate the class label of the test input by assigning the label of the nearest neighbor.

3.3 Unsupervised Learning Technique

In the intrusion detection based on supervised learning technique, we do not have any knowledge on the class labels of the input dataset. In such a situation, we aim to choose the classifier based unsupervised learning. This process is also called as clustering process. In unsupervised learning based technique the objective is to obtain a disjoint set of groups consisting of similar input objects. These groups may be used to perform decision making, to predict the future inputs.

The K-means clustering method is the most popular among the various clustering algorithms where k indicates the number of clusters to be formed from the input dataset. The K-means algorithm requires specifying the number of clusters to be formed well ahead.

In [20] the authors make use of this property to decide the number of clusters in their approach for intrusion detection.

3.4 Genetic Programming Approach

GA techniques are generally used to select best features that are used for IDS and when compared to other methods have a better efficiency. This approach is slightly tricky and complex and hence need to be used in specific manner rather in general approach.

Studies based on both Genetic Algorithms and Fuzzy rule based systems can be classified as Michigan, Pittsburgh approaches. The Pittsburgh method uses a set of if-then fuzzy rules which are coded as individual. In Michigan only a single if-then fuzzy rule is coded as individual.

4. Research Issues

The computation problems may be classified into two types. These include 1. Optimization Problems and 2. Decision problems. In optimization based problems, the objective is to aim for efficiency. In the decision based problems, we must output the decision as yes or no, true or false, etc. Intrusion detection may be considered as the decision problem where we need to classify if the target is an intrusion or not.

One approach of intrusion detection which is recently being concentrated is using text mining techniques. Data mining is a knowledge discovery process aiming at retrieving the unknown hidden information available, but not yet been identified and focused to derive important conclusions and findings. Intrusion detection and data mining have been complementing each other in research works performed by various researchers towards finding various possibilities, approaches to detect intrusion.

The data mining approaches such as prediction, classification, clustering, noise elimination have been extensively used in the intrusion detection process as discussed in the related works of section 2. In this section, our objective is to outline a generalized method for intrusion detection. The problem of predicting intrusion detection has been a major challenge for researchers from the medical domain as well as from the other fields of engineering such as health informatics, medical informatics and information retrieval. We now try to point out the various research issues in handling data sets.

4.1 Pre-processing Datasets

The research should first start with the study of the benchmark datasets. Sometimes there may be a need to start collecting data from scratch if we are working over a problem in a particular domain. Preprocessing phase is an essential phase to make the dataset suitable for process effectively to obtain accurate, efficient results by applying the newly designed method or already existing algorithm.

Since there is no specific standard dataset for intrusion detection, we choose to consider the KDD-Cup 99 dataset as the one considered in [20]. This contains 494,020 samples totally. The dimensionality of each data sample is 41. Of these 41 dimensions, a total of 9 are intrinsic type, 13 are content type while all the remaining 19 are traffic type. Each data sample of the dataset is classified in to 5 classes.

There are four types of attacks and one normal traffic class. Since the number of classes is five, this is basically as 5-Class Classification problem. Similarly we may choose to use the DARPA 1998 or 1999 standard dataset. In particular, the research should first start with the studying the benchmark datasets.

Sometimes there may be a need to start collecting data from scratch if we are working over a problem in particular domain. Preprocessing phase is an essential phase to make the dataset suitable for processing and handling effectively to obtain accurate, efficient results by applying the newly designed method or already existing algorithm.

4.2 Dimensionality Reduction

This phase includes extracting features from the dataset. These include feature selection, feature extraction, information gain, the application of linear discriminant analysis, noise elimination, dimensionality reduction by computing frequent patterns, etc. some of the recent works include application of text processing approaches for intrusion detection.

4.3 Distance Measure

The choice of distance measure is an essential task in prediction and classification processes. Some distance measures use the notation of vectors and other distance measures use non-vectors as input.

Some of the well-known distance measures include cosine distance measure, Manhattan distance, Euclidean distance, Jaccard measure. If the input is a frequency vector, we may use cosine measure for finding distance between the same. Alternately we may design our own measure to compute the distance between any two input entities.

4.4 Classification and Prediction Algorithm

The underlying dataset is the deciding factor for the choice of the algorithm. A single classification algorithm is not suitable for every dataset. Choosing an efficient classification method followed by inefficient distance measure may lead to improper estimation of intrusion prediction.

The existing classification algorithms have their own advantages and disadvantages, which need to be studied and chosen effectively.

4.5 Noise Elimination

In text mining based intrusion detection, we may have to form the process vs system call matrix for intrusion detection after finding the system call vector which contains all system calls. Since the dimensionality of the system call vector makes the dimensionality of system call matrix large, we may have to reduce the dimensionality.

After deciding the number of system calls, there may be one or more system calls which may be not important and may be discarded without any loss of information. Every effort must be made in this direction, so that the system call attributes which are of the least importance and insignificant affect may be eliminated.

5. Proposed Approach

The approach of intrusion detection using the text processing has been one of the research interests among researchers working in the area of network and information security. The proposed approach for intrusion detection is based on the concept of text processing and use of data mining techniques in the prediction and classification of intrusion.

Our intrusion detection is based on system calls. Formally, we treat the algorithm to be a function of system calls. In this approach for intrusion detection, the system calls serve as an important source for mining and predicting any chance of intrusion. When an application runs, there might be several system calls which are initiated in the background these system calls form the basis and the deciding factor for intrusion detection

5.1 Steps involved

The block schematic of the proposed approach is given in the figure 2 below. The following are the sequence of steps

5.1.1 Stage 1

The DARPA Dataset is used, as it is publicly available, labeled and pre-processed ready for use. Preprocessing of the dataset is to make it suitable for use by the data mining algorithm and techniques used to handle the data.

5.1.2 Stage 2

Perform dimensionality reduction of system calls, as all the system calls need not be important. We must identify those system calls which are not dominant and eliminate such system calls. The Singular Value Decomposition (SVD) technique may be used to perform the dimensionality reduction at this stage. By applying SVD, we can figure out most dominant and least dominant system calls. Such system calls which do not make any significant effect may be eliminated. All such system calls are called outliers. A simple thumb rule is to eliminate all the systems calls whose Eigen values are less than 1.

5.1.3 Stage 3

This stage involves deciding which system calls must be retained from the most dominant system calls obtained in the previous stage. A simple thumb rule is to consider all those system calls, which add up to 90% energy.

5.1.4 Stage 4

We may apply frequent pattern approaches for finding frequent system calls. In this case, we are trying to find the system call item sets. Finding system calls sets may be used to derive important association rules which may be helpful in performing classification and for predicting the trends among system calls. There is a scope for research in this direction as very less work is carried out.

5.1.5 Stage 5

This stage involves using suitable distance measure such Euclidean, Cosine, etc, Alternately, one may design his/her own kernel measure which may be used to perform classification. Such a distance measure which is designed must satisfy all the basic properties of the distance function [27].

5.1.6 Stage 5

The next stage involves the process of classification. There are less number of options available for the datasets. This process becomes much simpler as the DARPA dataset is used for this purpose. We may use DARPA dataset, as it is publicly available, preprocessed and ready for use. To perform this process there are two approaches, kernel based and distance based measures. Similarity measures such as the Cosine measure and the Jaccard measure such as various distance measures may be used. In the case of binary matrix representation of the system calls, the Jaccard distance measure may be used. The Cosine measure is used for the frequency based system calls. On the other hand, we may also use methods such as SVM classification.

Alternatively, the user may design a new kernel measure and use with SVM classifier to perform classification.

5.1.7 Research Direction

There is scope for research, if we make use of association rules to perform dimensionality reduction. Efforts are countable in this direction as very less work is performed by researchers. One way is to find the relation between the system calls and reduce the system calls which are not important, if we already know the class as intrusion and non-intrusion.

6. Text Mining Based Intrusion Detection

The consensus based computing approach has been applied in various application areas which aims at using more than one algorithm or procedure, distance measures to address the respective problems. Since the chosen dataset has already defined the number of classes, and the intrusion detection is also a classification problem, we may choose to cluster the chosen dataset into number of clusters equal to the number of class labels.

In this paper, the objective is to use the K-means clustering method to cluster the chosen dataset into a number of clusters equal to the number of class labels. We may directly cluster the training set or alternatively choose perform feature selection followed by dimensionality reduction and then apply K-means clustering over this reduced dimensionality.

6.1 Handling Training Set

We follow the approach in [20] for dimensionality reduction. However, instead of using the conventional k-means algorithm, we choose to apply the modified K-means algorithm which uses the Gaussian based distance measure to find the similarity between data samples when forming the clusters. This is where the novelty of our approach starts with. In this approach, we reduce the dimensionality of training set by first applying a suitable clustering to a number of clusters equal to number of known class labels. Since the intrusions datasets are have labeled attacks, we can decide the number of clusters to be obtained. The better choice is k-means clustering algorithm as it clusters the input to the predefined number of clusters.

After, obtaining the clusters the next step is to find the distance between each training data sample and all the cluster centers. This is the first distance value computed. In addition to this for every data sample with in a cluster, we find its nearest neighbor within that cluster by selecting the pair of minimum distance. This is the second distance value.

The two distances are added to get a single distance. Now each data sample is mapped to a single distance value instead of data sample expressed as a function of system call attributes when performing text mining based intrusion detection. For example, if we consider the purpose of clustering, we must specify the number of clusters equal

6.2 Distance Measure for K-means

In this section, we discuss the distance measure used as part of the k-means clustering algorithm. We use the Gaussian function as the distance measure to find the distance between any two samples of training set. This may also be used to find the distance between any two data samples in general.

6.2.1 Gaussian Function

We consider the Gaussian function based distance measure to find the similarity between the data samples of the intrusion dataset. We use the same distance measure and apply the k-means algorithm to cluster the data samples.

For the purpose of dimensionality reduction, we use the k-means clustering technique to obtain the clusters using the proposed distance function and then find the distance between each training data sample and each of the cluster centroids. This is further followed by finding the nearest neighbor for every data sample within the cluster. These two distances are summed to get a new distance value. This distance value becomes singleton feature for each training data sample. Thus each data sample of the training set is mapped to a single feature value reducing the dimensionality to 1.

The Proposed distance function is defined as given by Equation. 1

$$G(\, x, \mathrm{\mu}, \sigma) = \begin{cases} e^{-(\frac{x-\mathrm{\mu}}{\sigma})^2} \;; & one\ or\ both\ system\ calls\ exist \\ 0 \;; & none\ of\ the\ system\ calls\ exist \end{cases} \tag{1}$$

where
x = system call being considered
μ = mean of the system call w.r.t data samples present in the cluster
σ = standard deviation of system call considered w.r.t data samples of training set.

The denominator of IDSIM is given by Equation.2 as shown below

$$H(\, x, \mu, \sigma) = \begin{cases} 1 \;; & one\ or\ both\ system\ calls\ exist \\ 0 \;; & none\ of\ the\ system\ calls\ exist \end{cases} \tag{2}$$

The average distance is the ratio of the two functions $G(x, \mu, \sigma)$ *and* $H(x, \mu, \sigma)$ and is formally represented as given by Eq.3

$$F_{avg} = \frac{G(x,\mu,\sigma)}{H(x,\mu,\sigma)} \qquad (3)$$

The average distance considering distribution of all features hence is defined as Equation.4 as given below

$$F_{avg} = \frac{\sum_{i=1}^{i=n} 1 \sum_{s=1}^{s=m} e^{-(\frac{x_{is}-\mu_{is}}{\sigma_s})^2}}{\sum_{i=1}^{i=n} 1 \sum_{s=1}^{s=m} 1} \qquad (4)$$

The distance function is represented as given by

$$IDSIM = (1+F_{avg}) / 2 \qquad (5)$$

Where i indicates the ith data sample. S indicates the system call. IDSIM indicates the similarity function

6.3 Dimensionality Reduction of Training Set

Figure.3 shows the proposed approach for reducing the dimensionality of the training set and Figure.4 shows the proposed approach for reducing the dimensionality of the testing set using the proposed measure with K-means clustering technique.

So, we have both the testing and training sets with each data sample transformed to a singleton feature value. The test dataset can now be compared with training dataset in a very simple and effective, efficient way. The Proposed approach concentrates on using the Gaussian function based distance along with the K-means instead of conventional distance function used by K-means algorithm.

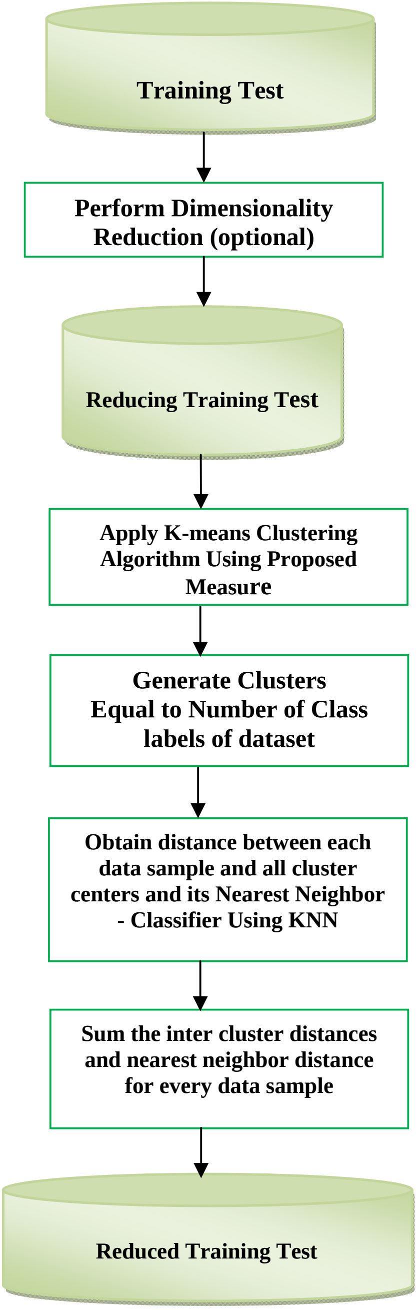

Training Test
Perform Dimensionality Reduction (optional)
Reducing Training Test
Apply K-means Clustering Algorithm Using Proposed Measure
Generate Clusters Equal to Number of Class labels of dataset
Obtain distance between each data sample and all cluster centers and its Nearest Neighbor - Classifier Using KNN
Sum the inter cluster distances and nearest neighbor distance for every data sample
Reduced Training Test

Fig 3: Dimensionality Reduction of Training Set

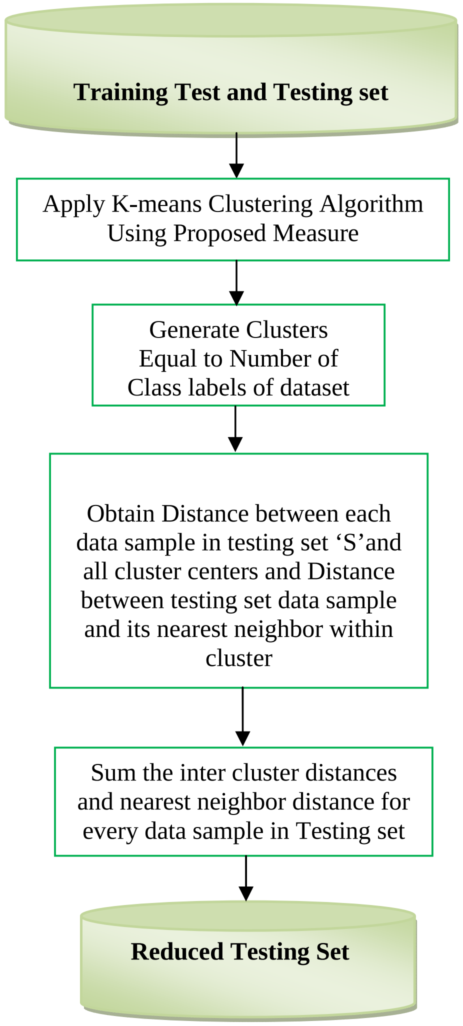

Training Test and Testing set
Apply K-means Clustering Algorithm Using Proposed Measure
Generate Clusters Equal to Number of Class labels of dataset
Obtain Distance between each data sample in testing set 'S'and all cluster centers and Distance between testing set data sample and its nearest neighbor within cluster
Sum the inter cluster distances and nearest neighbor distance for every data sample in Testing set
Reduced Testing Set

Fig 4: Dimensionality Reduction of Testing set

7. Conclusions

This paper discusses various approaches to be followed for detection of the intrusion. It also discusses the research issues to be considered in intrusion detection using text processing. It also discusses the sequence of steps to be followed in order to improve the effectiveness and performance of the intrusion detection mechanism and decreasing false alarms that are generated by the intrusion detection system. This paper also discusses the importance of preprocessing techniques, dimensionality reduction in order to reduce the false alarms. In this work, the second major contribution is in defining the similarity measure which has finite lower and upper bounds. The measure designed is Gaussian function based distance measure. The K-means algorithm is chosen for clustering using the proposed distance measure to cluster both the training and testing data samples. The training and test datasets are transformed to single dimensional feature with the use of k-means and proposed distance measure. The significance of the proposed distance measure is it considers the distribution of the system calls behavior over the entire training samples. This makes the computation accurate, even in binary form. The similarity value lies between 0 and 1.

References

[1] Alok Sharma, Arun K. Pujari, and Kuldip K. Paliwal. 2007. Intrusion detection using text processing techniques with a kernel based similarity measure. Comput. Secur. 26, 7-8 (December 2007), 488-495.

[2] Yihua Liao and V. Rao Vemuri. 2002. Using Text Categorization Techniques for Intrusion Detection. In Proceedings of the 11th USENIX Security Symposium, Dan Boneh (Ed.). USENIX Association, Berkeley, CA, USA, 51-59.

[3] Sanjay Rawat, Arun K. Pujari, and V. P. Gulati. 2006. On the Use of Singular Value Decomposition for a Fast Intrusion Detection System. Electron. Notes Theor. Comput. Sci. 142 (January 2006), 215-228.

[4] Sang Hyun Oh and Won Suk Lee. 2003. Refereed papers: An anomaly intrusion detection method by clustering normal user behavior. Comput. Secur. 22, 7 (October 2003), 596-612.

[5] Sang Hyun Oh and Won Suk Lee. 2003. Refereed papers: An anomaly intrusion detection method by clustering normal user behavior. Comput. Secur. 22, 7 (October 2003), 596-612.

[6] Arman Tajbakhsh, Mohammad Rahmati, and Abdolreza Mirzaei. 2009. Intrusion detection using fuzzy association rules. Appl. Soft Comput. 9, 2 (March 2009), 462-469.

[7] Wenke Lee; Stolfo, S.J.; Mok, K.W., ”A data mining framework for building intrusion detection models,” in Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium on , vol., no., pp.120-132, 1999

[8] P. Garca-Teodoro, J. Daz-Verdejo, G. Maci-Fernndez, and E. Vzquez. 2009. Anomaly-based network intrusion detection: Techniques, systems and challenges. Comput. Secur. 28, 1-2 (February 2009), 18-28.

[9] Bridges S.M., Vaughn R.B. Fuzzy data mining and genetic algorithms applied to intrusion detection. In: Proceedings of the National Information Systems Security Conference; 2000. pp. 1331.

[10] Dickerson, J.E.; Dickerson, J.A., ”Fuzzy network pro- filing for intrusion detection,” in Fuzzy Information Processing Society, 2000. NAFIPS. 19th International Conference of the North American , vol., no., pp.301- 306, 2000

[11] Vladimir N. Vapnik. 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., New York, NY, USA.

[12] Thorsten Joachims. 1999. Making large-scale support vector machine learning practical. In Advances in kernel methods, Bernhard Schlkopf, Christopher J. C. Burges, and Alexander J. Smola (Eds.). MIT Press, Cambridge, MA, USA 169-184.

[13] Neminath Hubballi and Vinoth Suryanarayanan. 2014. Review: False alarm minimization techniques in signature-based intrusion detection systems: A survey. Comput. Commun. 49 (August 2014), 1-17.

[14] Yuxin Meng, Wenjuan Li, and Lam-For Kwok. 2013. Towards adaptive character frequency-based exclusive signature matching scheme and its applications in distributed intrusion detection. Comput. Netw. 57, 17 (December 2013), 3630-3640.

[15] Wenke Lee and Salvatore J. Stolfo. 1998. Data mining approaches for intrusion detection. In Proceedings of the 7th conference on USENIX Security Symposium - Volume 7 (SSYM’98), Vol. 7. USENIX Association, Berkeley, CA, USA, 6-6.

[16] Daniel Barbará, Julia Couto, Sushil Jajodia, and Ningning Wu. 2001. ADAM: a testbed for exploring the use of data mining in intrusion detection. SIGMOD Rec. 30, 4 (December 2001), 15-24.

[17] Lee, W. Stolfo, S. Kui, M.: A Data Mining Framework for Building Intrusion Detection Models. IEEE Symposium on Security and Privacy (1999) 120-132

[18] Manganaris, S., Christensen, M., Zerkle, D., Hermiz, K.: Stefanos Manganaris, Marvin Christensen, Dan Zerkle, and Keith Hermiz. 2000. A data mining analysis of RTID alarms. Comput. Netw. 34, 4 (October 2000), 571-577.

[19] James J. Treinen and Ramakrishna Thurimella. 2006. A framework for the application of association rule mining in large intrusion detection infrastructures. In Proceedings of the 9th international conference on Recent Advances in Intrusion Detection (RAID’06), Diego Zamboni and Christopher Kruegel (Eds.). SpringerVerlag, Berlin, Heidelberg, 1-18.

[20] M. Govindarajan and RM. Chandrasekaran. 2011. Intrusion detection using neural based hybrid classification methods. Comput. Netw. 55, 8 (June 2011), 1662- 1671.

[21] Richard P. Lippmann and Robert K. Cunningham. 2000. Improving intrusion detection performance using keyword selection and neural networks. Comput. Netw. 34, 4 (October 2000), 597-603.

[22] Chirag Modi, Dhiren Patel, Bhavesh Borisaniya, Hiren Patel, Avi Patel, and Muttukrishnan Rajarajan. 2013. Review: A survey of intrusion detection techniques in Cloud. J. Netw. Comput. Appl. 36, 1 (January 2013), 42-57.

[23] Mohammad Saniee Abadeh, Hamid Mohamadi, and Jafar Habibi. 2011. Design and analysis of genetic fuzzy systems for intrusion detection in computer networks. Expert Syst. Appl. 38, 6 (June 2011), 7067-7075.

[24] Lee, W., Salvatore, J. S., Mok, K. W. M. (1998). Mining audit data to build intrusion

detection Models. In Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (pp. 66-72).

[25] Chirag Modi, Dhiren Patel, Bhavesh Borisaniya, Hiren Patel, Avi Patel, and Muttukrishnan Rajarajan. 2013. Review: A survey of intrusion detection techniques in Cloud. J. Netw. Comput. Appl. 36, 1 (January 2013), 42-57.

[26] Sanjay Rawat, V P Gulati, Arun K Pujari, ”A Fast Host-Based Intrusion Detection System Using Rough Set Theory”, J. Springer Transactions on Rough Sets IV, Lecture Notes in Computer Science Volume 3700, 2005, pp 144-161

[27] Yung-Shen Lin; Jung-Yi Jiang; Shie-Jue Lee, ”A Similarity Measure for Text Classification and Clustering,” in Knowledge and Data Engineering, IEEE Transactions on , vol.26, no.7, pp.1575-1590, July 2014

IJCSIS REVIEWERS’ LIST

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2016
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on

its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC’s, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/